

格致方法 · 定量研究系列　吴晓刚　主编

定序题项回答理论：莫坎量表分析

[荷] 韦杰勃朗 · H. 凡舒尔（Wijbrandt H. van Schuur）著
王佳 译　缪佳 校

革新研究理念

丰富研究工具

最权威、最前沿的定量研究方法指南

64

格致出版社　上海人民出版社



针对社会科学中的测量问题，本书简明扼要地介绍了如何测量二分和定序问题的定序题项回答理论（IRT）、模型以及在量表中的应用。

本书可以帮助那些对测量个体潜在特性感兴趣的社会科学工作者更好地理解定序量表背后的假设，在不同的假设下如何选择量表题项和构建量表。作者穿插使用了大量的概念和量表测量实例，生动形象地展示了题项回答理论在不同模型假设下的实际应用。

主要特点

- 本书可以帮助社会科学工作者更深刻地思考测量背后个人和量表题项回答的关系，而不仅仅简单假定理想测量的存在
- 题项回答理论（IRT）和模型可以提供更多实际量表来拟合社会和行为科学中不同种类的数据
- 作者同时比较了定序 IRT 模型与其他测量模型在分析结果上可能存在的差异及背后的原因

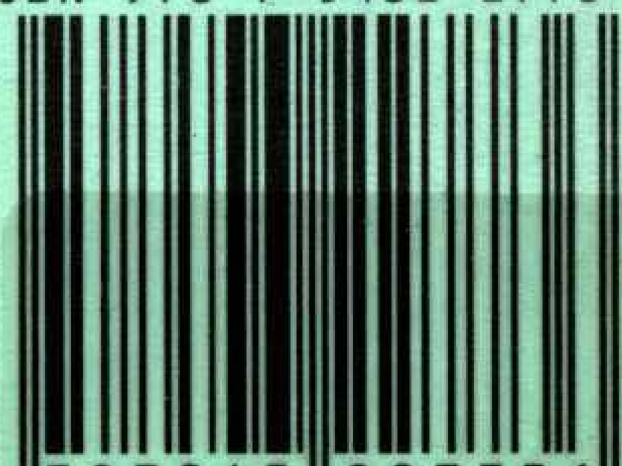
上架建议：社会研究方法

您可以通过如下方式联系到我们：
邮箱：hibooks@hibooks.cn


微信


天猫

ISBN 978-7-5432-2773-6



定价：32.00元
易文网：www.ewen.co
格致网：www.hibooks.cn

格致方法·定量研究系列　吴晓刚　主编

定序题项回答理论：莫坎量表分析

[荷]韦杰勃朗·H.凡舒尔（Wijbrandt H.van Schuur）
王　佳 译　缪　佳 校

SAGE Publications, Inc.
格致出版社　上海人民出版社

图书在版编目(CIP)数据

定序题项回答理论:莫坎量表分析/(荷)韦杰勃朗・H.凡舒尔著;王佳译;缪佳校.—上海:格致出版社:上海人民出版社,2017.9
(格致方法・定量研究系列)
ISBN 978-7-5432-2773-6

Ⅰ.①定… Ⅱ.①韦… ②王… ③缪… Ⅲ.①问答法-评定量表 Ⅳ.①G424.1

中国版本图书馆CIP数据核字(2017)第161887号

责任编辑 贺俊逸

格致方法・定量研究系列
定序题项回答理论:莫坎量表分析
[荷]韦杰勃朗・H.凡舒尔 著
王 佳 译 缪 佳 校

出 版	世纪出版股份有限公司 格致出版社 世纪出版集团 上海人民出版社 (200001 上海福建中路193号 www.ewen.co) 编辑部热线 021-63914988 市场部热线 021-63914081 www.hibooks.cn	**印 刷**	浙江临安曙光印务有限公司
		开 本	920×1168 1/32
		印 张	6.25
		字 数	125,000
		版 次	2017年9月第1版
发 行	上海世纪出版股份有限公司发行中心	**印 次**	2017年9月第1次印刷

ISBN 978-7-5432-2773-6/C・184 定价:32.00元

出版说明

由香港科技大学社会科学部吴晓刚教授主编的“格致方法·定量研究系列”丛书，精选了世界著名的SAGE出版社定量社会科学研究丛书，翻译成中文，起初集结成八册，于2011年出版。这套丛书自出版以来，受到广大读者特别是年轻一代社会科学工作者的热烈欢迎。为了给广大读者提供更多的方便和选择，该丛书经过修订和校正，于2012年以单行本的形式再次出版发行，共37本。我们衷心感谢广大读者的支持和建议。

随着与SAGE出版社合作的进一步深化，我们又从丛书中精选了三十多个品种，译成中文，以飨读者。丛书新增品种涵盖了更多的定量研究方法。我们希望本丛书单行本的继续出版能为推动国内社会科学定量研究的教学和研究作出一点贡献。

总序

2003年，我赴港工作，在香港科技大学社会科学部教授研究生的两门核心定量方法课程。香港科技大学社会科学部自创建以来，非常重视社会科学研究方法论的训练。我开设的第一门课"社会科学里的统计学"(Statistics for Social Science)为所有研究型硕士生和博士生的必修课，而第二门课"社会科学中的定量分析"为博士生的必修课(事实上，大部分硕士生在修完第一门课后都会继续选修第二门课)。我在讲授这两门课的时候，根据社会科学研究生的数理基础比较薄弱的特点，尽量避免复杂的数学公式推导，而用具体的例子，结合语言和图形，帮助学生理解统计的基本概念和模型。课程的重点放在如何应用定量分析模型研究社会实际问题上，即社会研究者主要为定量统计方法的"消费者"而非"生产者"。作为"消费者"，学完这些课程后，我们一方面能够读懂、欣赏和评价别人在同行评议的刊物上发表的定量研究的文章；另一方面，也能在自己的研究中运用这些成熟的方法论技术。

上述两门课的内容，尽管在线性回归模型的内容上有少

量重复,但各有侧重。“社会科学里的统计学”从介绍最基本的社会研究方法论和统计学原理开始,到多元线性回归模型结束,内容涵盖了描述性统计的基本方法、统计推论的原理、假设检验、列联表分析、方差和协方差分析、简单线性回归模型、多元线性回归模型,以及线性回归模型的假设和模型诊断。“社会科学中的定量分析”则介绍在经典线性回归模型的假设不成立的情况下的一些模型和方法,将重点放在因变量为定类数据的分析模型上,包括两分类的 logistic 回归模型、多分类 logistic 回归模型、定序 logistic 回归模型、条件 logistic 回归模型、多维列联表的对数线性和对数乘积模型、有关删节数据的模型、纵贯数据的分析模型,包括追踪研究和事件史的分析方法。这些模型在社会科学研究中有着更加广泛的应用。

修读过这些课程的香港科技大学的研究生,一直鼓励和支持我将两门课的讲稿结集出版,并帮助我将原来的英文课程讲稿译成了中文。但是,由于种种原因,这两本书拖了多年还没有完成。世界著名的出版社 SAGE 的“定量社会科学研究”丛书闻名遐迩,每本书都写得通俗易懂,与我的教学理念是相通的。当格致出版社向我提出从这套丛书中精选一批翻译,以飨中文读者时,我非常支持这个想法,因为这从某种程度上弥补了我的教科书未能出版的遗憾。

翻译是一件吃力不讨好的事。不但要有对中英文两种语言的精准把握能力,还要有对实质内容有较深的理解能力,而这套丛书涵盖的又恰恰是社会科学中技术性非常强的内容,只有语言能力是远远不能胜任的。在短短的一年时间里,我们组织了来自中国内地及香港、台湾地区的二十几位

研究生参与了这项工程，他们当时大部分是香港科技大学的硕士和博士研究生，受过严格的社会科学统计方法的训练，也有来自美国等地对定量研究感兴趣的博士研究生。他们是香港科技大学社会科学部博士研究生蒋勤、李骏、盛智明、叶华、张卓妮、郑冰岛，硕士研究生贺光烨、李兰、林毓玲、肖东亮、辛济云、於嘉、余珊珊，应用社会经济研究中心研究员李俊秀；香港大学教育学院博士研究生洪岩璧；北京大学社会学系博士研究生李丁、赵亮员；中国人民大学人口学系讲师巫锡炜；中国台湾“中央”研究院社会学所助理研究员林宗弘；南京师范大学心理学系副教授陈陈；美国北卡罗来纳大学教堂山分校社会学系博士候选人姜念涛；美国加州大学洛杉矶分校社会学系博士研究生宋曦；哈佛大学社会学系博士研究生郭茂灿和周韵。

参与这项工作的许多译者目前都已经毕业，大多成为中国内地以及香港、台湾等地区高校和研究机构定量社会科学方法教学和研究的骨干。不少译者反映，翻译工作本身也是他们学习相关定量方法的有效途径。鉴于此，当格致出版社和 SAGE 出版社决定在“格致方法 · 定量研究系列”丛书中推出另外一批新品种时，香港科技大学社会科学部的研究生仍然是主要力量。特别值得一提的是，香港科技大学应用社会经济研究中心与上海大学社会学院自 2012 年夏季开始，在上海（夏季）和广州南沙（冬季）联合举办《应用社会科学研究方法研修班》，至今已经成功举办三届。研修课程设计体现“化整为零、循序渐进、中文教学、学以致用”的方针，吸引了一大批有志于从事定量社会科学研究的博士生和青年学者。他们中的不少人也参与了翻译和校对的工作。他们在

繁忙的学习和研究之余,历经近两年的时间,完成了三十多本新书的翻译任务,使得"格致方法·定量研究系列"丛书更加丰富和完善。他们是:东南大学社会学系副教授洪岩璧,香港科技大学社会科学部博士研究生贺光烨、李忠路、王佳、王彦蓉、许多多,硕士研究生范新光、缪佳、武玲蔚、臧晓露、曾东林,原硕士研究生李兰,密歇根大学社会学系博士研究生王骁,纽约大学社会学系博士研究生温芳琪,牛津大学社会学系研究生周穆之,上海大学社会学院博士研究生陈伟等。

陈伟、范新光、贺光烨、洪岩璧、李忠路、缪佳、王佳、武玲蔚、许多多、曾东林、周穆之,以及香港科技大学社会科学部硕士研究生陈佳莹,上海大学社会学院硕士研究生梁海祥还协助主编做了大量的审校工作。格致出版社编辑高璇不遗余力地推动本丛书的继续出版,并且在这个过程中表现出极大的耐心和高度的专业精神。对他们付出的劳动,我在此致以诚挚的谢意。当然,每本书因本身内容和译者的行文风格有所差异,校对未免挂一漏万,术语的标准译法方面还有很大的改进空间。我们欢迎广大读者提出建设性的批评和建议,以便再版时修订。

我们希望本丛书的持续出版,能为进一步提升国内社会科学定量教学和研究水平作出一点贡献。

吴晓刚

于香港九龙清水湾

序

在心理物理学家沃伦·托格森(Warren Torgerson)1958年的里程碑著作《量表理论和方法》(*Theory and Methods of Scaling*)一书中,他区分出了基本测量以及指定测量。指定测量为测量对象主观指定数字,这种指定通常是根据合理性或者常识来做出的。这种测量最常见的例子就是里克特或者评分加总量表,其中连续的整数被赋予定序的类别:比如1代表强烈不同意,2代表不同意,3代表同意,以及4代表强烈同意。若干条题项的数字加总起来就构成了一个量表。虽然伦西斯·里克特(Rensis Likert)在介绍这一方法时确实试图证明:赋予分类以不同数字也会产生高度相关的量表,但是他并没有进一步论证:为什么要采取各类别之间是等距的,或者为什么各项分数可以直接加总。评分加总量表的特征,比如信度,是经典测试理论的研究内容。

托格森的著作所关心的内容不同于心理学和社会科学的基本测量方法,它包含的是稍后被纳入题项回答理论中的方法。在基本测量中赋予对象的数字都必须遵循一定的规则。举例来说,在一个定序量表中,如果一个对象被赋予了

比另一个对象更大的数字,那么必须证明其具有更多的测量对象的特性。同样地,要论证一个等距量表,有必要证明量表数值之间的相同差异实际反映了测量对象在数量上的相同差异。另一种表述方法是,基本测量模型关于测量结构有很强的假定,这些假定可以用数据来证伪。

最简单的题项回答模型涉及二分回答问题,用克莱德·库姆(Clyde Coomb)1964 年的开创性著作《数据理论》(*A Theory of Data*)中的术语来讲,个人与题项处于一种支配关系。比如考试中的问题。个人是否正确答对考试题目同时取决于个人的知识或能力以及题目的难度。在这样的模型中个人和题项被假定同时存在于一个共同的能力和难度的维度之上。比如,在确定性的哥特曼量表(deterministic Guttman-scale)模型中,个人的能力超过问题的难度时就必然会做出正确回答,否则就做出错误回答。概率题项回答模型,比如托格森讨论的正态肩形模型,以及 20 世纪 60 年代由丹麦心理测量学家格奥尔格·拉希(Georg Rasch)引入的相似的logistic模型,则更加细微:正确回答某一题项的概率作为个人能力与题项难度的差异的函数而平稳增加。

从哥特曼、托格森、拉希和库姆的时代以来,题项回答理论发展迅速。虽然韦杰勃朗·H.凡舒尔(Wijbrandt H. van Schuur)在本书中讨论了一系列题项回答模型,包括哥特曼量表和拉希模型,本书的重点关注针对二分和定序问题的定序题项回答理论,准确地说,即传统上构成评分加总量表的数据类型。回答的概率是问题和个人在潜在量度上的位置的联合分布,这些模型不假定这种联合分布是正态的、logistic 的,或者其他任何一种特定分布。相反,题项回答函

数被简单假定为单调函数。这些模型首先由荷兰统计学家和社会科学家罗伯特·J.莫坎(Robert J.Mokken)引入,得到的是定序量表。

通过放宽题项回答曲线符合某种特定函数这一假设,定序题项回答理论提供了得到更多实际量表的可能性,以拟合社会和行为科学中更多种类的数据。凡舒尔在介绍背景、理论依据和定序问题理论的结构方面做了很好的工作。我期待他这本专著会推动研究者更多地应用这些模型,从而带来社会测量的改进。

约翰·福克斯

致谢

作者和SAGE出版社感谢下列评论者做出的贡献：

大卫·安德里克(David Andrich)，西澳大学

威廉·G.雅各布(William G. Jacoby)，密歇根州立大学和 *ICPSR*

献给埃里克·坦嫩鲍姆(Eric Tanenbaum)，社会科学数据分析和收集学院艾塞克斯暑期学校前主任。他一直保持着对最前沿的方法创新的敏锐把握。

目 录

序 1

致谢 1

第 1 章 概论 1

第 1 节 社会科学中的测量问题 2

第 2 节 测量理论以及二分题项的题项回答理论 4

第 3 节 两种基本不同的 IRT 模型 5

第 2 章 哥特曼量表 7

第 1 节 成对问题之间的特殊关系 8

第 2 节 使用问题的答案作为测量工具 13

第 3 节 多于两个问题的定序测量 14

第 4 节 哥特曼量表:理想的确定性累计量表 17

第 5 节 确定性模型的假设 20

第 6 节 路易斯·哥特曼和哥特曼量表 21

第 3 章 非理想的累计量表 23

第 1 节 模型违反 24

第 2 节　误差:违反题项和研究对象之间的传递关系　26
第 3 节　在较大数据集中误差定义的扩展　30
第 4 节　如何评价数据集中的模型违反数量　32
第 5 节　评价同质性系数　38
第 6 节　在具有超过两个题项的量表中使用同质性系数　40
第 7 节　误差的“原因”:题项还是研究对象?　45
第 8 节　“归咎于”研究对象:转置数据矩阵及计算研究对象同质性　47
第 9 节　使用非理想模式来测量研究对象量表值　49
第 10 节　结论　50

第 4 章　确定或寻找构成量表的题项　51

第 1 节　寻找可以构成量表的题项　52
第 2 节　不在量表中的题项:拒绝和排除　56

第 5 章　一个累计性量表的例子:美国宗教信念　59

第 1 节　寻找相关的基础信息　62
第 2 节　总结必要的基本信息　64
第 3 节　计算单个题项和整个量表的同质性　65
第 4 节　统计显著性检验　66
第 5 节　使用成对的信息寻找最佳的量表　68
第 6 节　使用转置的二分数据矩阵　70
第 7 节　使用新建立的量表的参数　72

第 8 节 结论 74

第 6 章 概率性支配模型：单调同质性 75

第 1 节 不理想的回答还是概率性的回答？ 76

第 2 节 两个概率性模型:单调同质性和双重单调性 80

第 3 节 检验单调同质性模型 83

第 4 节 检验五个宗教信念题项的单调同质性 88

第 7 章 概率性支配模型：双重单调性 91

第 1 节 双重单调性的重要意义 92

第 2 节 使用外部群体测试双重单调性 93

第 3 节 使用余分分组测试双重单调性 98

第 4 节 使用合并余分分组测试双重单调性 102

第 5 节 使用 P(＋，＋)和 P(－，－)矩阵测试双重单调性 104

第 6 节 从概率性模型的测试中我们可以学到什么？ 109

第 8 章 多分类题项的累计测量 111

第 1 节 多分类题项构成的确定性累计模型的回答模式 114

第 2 节 在确定性累计量表中使用社会科学题项 117

第 3 节 评价同质性 122

第 4 节 多分类题项的寻找程序 129

第 5 节 对多分类题项应用概率性模型 131
第 6 节 政治行为量表 134

第 9 章 余论 137
第 1 节 信度分析 139
第 2 节 因子分析 143
第 3 节 参数 IRT 模型:拉希模型 145
第 4 节 对美国人宗教信念数据应用其他测量模型 147
第 5 节 不具区分度的回答模式 149
第 6 节 一些实际问题 151
第 7 节 一些最后的评论 153

附录 155
注释 162
参考文献 165
进一步阅读书目 168
译名对照表 172

第1章

概　论

第1节 | 社会科学中的测量问题

在我们的日常交谈中,我们经常提及那些人们很难观察到的特征。我们可能会形容某个人“他非常活跃”“她非常机智”,或者“他比他妻子更加保守”。这些结论是我们在观察到某一特定情形下的活动、表现出来的智慧或者保守后推断出来的。但是我们不仅仅满足于此;我们还会将这些结论一般化。我们假定这些观察告诉了我们人们在这些特定情形之外的特征或者性格。当前表现得积极、智慧或者保守的个人在其他方面也将会(我们相信)继续采用这种方式行动,总体上他们会(我们相信)比其他人更加积极、更加智慧,或者更加保守。社会科学家想要测量这些个体的性格特质或者特征。由于这些特征只能够通过观察特定情境下的行为来推断,我们称之为潜在的特性。潜在的特性与外在的特性相对,后者是指那些可以直接通过观察并且(通常来说)不会改变的特征,如性别或者种族背景。

社会科学家通常对潜在的特性感兴趣,比如能力(如活动、知识,或者智力的水平)或者态度(如保守主义、信任,或者宗教信念)。他们相信这些能力和态度对于解释人们的行为方式以及行为决策很重要。这给社会科学家的工作带来了困难,因为人们的这些特征是很难测量到的。如果能够使

用简单的测量单位，例如沿用货币单位或者自然科学中的单位，如血压、大脑电位或者血液的化学组成，那么社会科学研究会变得更加容易。

另外一个问题是社会科学家想要测量的大部分潜在特征都没有已成体系的测量单位。举例来说，在测量异化时，我们不能用 milli-Marx 作为单位；测量挫折时，也不能以 kilo-Freud 为单位。社会科学家还没有能够建立起社会科学的显微镜或者望远镜。只有在他们愿意接受许多限制性假设的条件下，他们才能够使用数量的（或者基本水平的、定距水平的）测量。他们可能需要接受定序的测量，即根据人们的能力或者态度进行排序。在这种情形下我们把这种测量工具叫作*定序量表*，有时候也被称为非度量或者非参数量表。

第2节 测量理论以及二分题项的题项回答理论

我们可能会怀疑潜在特征比如能力和态度是否可能被测量。答案是肯定的，但是只有在我们准备好接受特定假设的条件下。这些假设在理论中被明确地表示出来。因为我们大部分的测量依赖于对测试或者调查问题的答案的解释，我们的假设与人们如何回答问题的理论有关。一种理论被称为题项回答理论(item response theory，IRT)。问题被称作题项，个体的行为(即对某一问题的答案)则被称为回答。最简单的行为是二分回答，只有两个选项。比如，一个人有或者没有做过某一行动；对于某一知识性问题她的答案是对的还是错的；对于一条保守性陈述他是同意还是不同意。IRT以一种数学(测量)模型的形式来呈现。这个模型并不像它听起来那么可怕。实际上，理解这一模型的每一步都非常简单。

首先，我们介绍只使用二分回答来测量个体潜在特征的模型。之后，我们再将测量模型进一步扩展到超过两种回答的情况。为了易于说明，我们使用具有以下两个回答类别的二分类题项：一个正向回答和一个负向回答。两者中的哪一个会被研究者定义为正向回答则取决于要测量的潜在特征的方向(例如，活跃与不活跃，有知识的与知识贫乏的，保守主义与自由主义)，并且没有道德上的意义。

第3节 | 两种基本不同的IRT模型

我们可以区分出人们为何对某一问题给出正向（“是”）回答的两类基本原因。让我用两组不同的问题来说明这一点，这些问题都可以使用“是”或“否”来回答。

1a. 请问你的身高大约是1.80米(5英尺11英寸)吗?	是/否
1b. 请问你的身高大约是1.70米(5英尺7英寸)吗?	是/否
2a. 请问你是不是至少有1.80米(5英尺11英寸)高?	是/否
2b. 请问你是不是至少有1.70米(5英尺7英寸)高?	是/否

所有的四个问题都包括了一个作为基准水平的高度。对于问题1a和1b，一个人只有在他的身高确实接近1.80米或1.70米的条件下才会给出正向的回答。个人的身高和问题中提到的高度是相同的。或者换一种稍微不同的说法，个人的身高与问题提到的高度之间的差异是可以忽略的，它们的相似度和接近度非常高。相反，负向回答则不那么明确：身高不近似1.70米的个人可能是比1.70米要高或者更矮；这两种可能性是相反的。

每个对问题2a或2b做出肯定回答的个人都是身高超过1.80米或1.70米的。而且对问题2a和2b做出肯定回答的人群存在重合，因为高过1.80米的人必然也高过1.70米。像这种“比……大”或者“比……高”的关系叫作支配关系。

前一段文字描述的近似关系，则叫作临近关系。

在身高这一变量上，所有的人都有一个数字：他们的数值，即他们的身高，是用厘米或者英尺和英寸来表示。但是——这一点可能刚开始比较难以理解——每一条问题或题项也可以用一个数字或数值来表示，即问题中提到的基准水平。问题 1a 和 1b 从具有相同(近似地)数字的个人那里得到正向回答。问题 2a 和 2b 从具有更大数字的个人那里得到正向回答。对问题 1a 和 1b 的正向回答叫作*临近*回答，对问题 2a 和 2b 的正向回答叫做*支配*回答。[1]

本章只涉及支配回答，以及处理它们的 IRT 模型：*支配模型*。它有时候也被称作哥特曼量表，是用它的创造者路易斯·哥特曼的名字命名的(Louis Guttman, 1950)。

支配模型有两种类型：度量的和定序的。差别在于，在度量模型中我们把个体和题项具有的数字看作具有定距尺度的数字，而在定序模型中，这些数字只表示等级次序，即这些数字都处于一个定序量表上。本书只介绍定序支配模型(关于度量模型，参见 Andrich, 1988; Bond & Fox, 2007; Embretson & Reise, 2000; Ostini & Nering, 2006; Smith & Stone, 2009)。我们会在第 9 章介绍定序模型对于在多元线性统计方法(比如回归分析或者方差分析)中使用量表得分有什么影响。

第2章

哥特曼量表

第1节 | 成对问题之间的特殊关系

为了理解哥特曼量表分析背后的原理,我们来比较一下几对问题之间的答案。每一对都有两个问题,我们把它们叫作问题 A 和问题 B。比如,

身高	
问题 A:你的身高超过 1.70 米了吗?	是/否
问题 B:你的身高超过 1.80 米了吗?	是/否
计算能力	
问题 A:学生答对 2+2 等于多少了吗?	是/否
问题 B:学生答对 23×17 等于多少了吗?	是/否
政治意识形态	
问题 A:收入差距应该被缩小吗?	是/否
问题 B:收入差距应该被消除吗?	是/否
宗教信念	
问题 A:你相信天堂吗?	是/否
问题 B:你相信地狱吗?	是/否
富裕程度	
问题 A:你有 CD 播放机吗?	是/否
问题 B:你有洗碗机吗?	是/否

这五对问题具有一个共同点。在每一对问题中,对问题 A 回答“是”的人数要比对问题 B 回答“是”的人数多。并

且——这一点很重要——在每对问题中,如果人们对问题B回答“是”,那么一般他们也会对问题A回答“是”。这一点只对于第一对问题是逻辑上成立的。如果个人身高超过1.80米,那么他们必然也比1.70米高。但是对于其他四对问题而言,A和B的这种关系并不总是成立。调查还显示:如果一个人对问题B回答“是”,那么他/她一般也会对问题A选择“是”。稍后我会讲到这种关系的例外情况,但是现在我们假定A和B的关系对于这四对问题也同样存在。

我们如何解释这一事实:同意问题B的人同时也同意问题A?一般的答案是,个体的回答背后存在一个更为一般意义上的变量,每对问题给出了个体在这一变量上的取值。第一对问题可以很清楚地说明这一点。如果个人超过1.80米,他们在身高变量上的测量数值,或者量表数值,就比1.80米要大。如果他们身高不超过1.70米,那么他们的身高数值就比1.70米小。并且如果他们比1.70米高但是不超过1.80米,那么他们的身高数值就介于1.70米和1.80米之间。

同样的原则也适用于第二对问题。我们认为问题的答案由一个潜在的变量决定,我们称之为计算能力。正确回答问题B比正确回答问题A需要的计算能力更高。因此这两个问题可以用来区分个人在潜在变量上处于下列三个类别中的哪一个(即,具有三个量表数值中的一个):

0:他们不能够正确回答两个问题中的任何一个。

1:他们只能够正确回答问题A。

2:他们能够正确回答两个问题。

这个例子与第一个例子的不同之处在于,测量人们的身高有直接的工具,然而对于测量计算能力则没有。我们不能将个人连接到某个机械或者电子测量设备上,通过刻度盘读数来获知他们的计算能力。如果我们想测量某人的计算能力,必须使用像问题A或者问题B那样的非直接指标。如果相对于问题A来说,问题B的难度更大,回答正确的可能性更小,我们就可以说对B的正确回答也暗示了对A的正确回答。具有此类问题的哥特曼量表因此有时也被称为蕴含量表。从这种意义上的蕴含可以得出正确回答问题A的人比正确回答问题B的人要多。在一对问题中,我们用较容易这一词语来形容更加大众的问题(即具有更多正向回答的问题),用较难这一词语来形容不那么大众的问题(即具有较少正向回答的问题)。所以在目前我们所有的例子中A是较容易的,B是较难的。

第三对问题测量的是政治观点。问题B表达的观点被认为比问题A表达的观点更加极端。同意任何一种说法都表明个人希望收入差距缩小。但是同意问题B的个人想要在这一点上做得更多。所以对这两个问题的回答把人们分成了三组:

0:认为收入差距不应该缩小的群体。

1:认为收入差距应该缩小但是不应该完全消除的群体。

2:认为收入差距应该被完全消除的群体。

这两个问题间接测量的变量经常被称为政治意识形态,或者更确切地说,社会经济方面左翼—右翼的维度划分,或

者自由—保守的维度划分。因为类似计算能力和政治意识形态这些变量不能直接观察到,它们经常被称为未观察到的变量,意思是它们不能够由任何一种直接测量工具观察到。这两个问题A和B被称为未观察到的变量的指标。

未观察到的变量的另一种名称叫作潜变量。潜在这一词语似乎意味着变量并不真正存在,只有当相关的指标问题被问到的时候它才会跳出来。如前所述,一些心理测量学家不讨论潜变量,而是讨论潜在的特性,特性指特定的人格特征。就哥特曼量表的应用来说,我们不必详细解释未观察到的变量。在大部分情况下,我们交替使用这些不同的名词,而不深究它们的心理测量学上的意义。

第四对问题涉及宗教信念。相信有地狱存在的人也相信天堂的存在,而并不是每一个相信天堂存在的人都相信地狱的存在。我们可以从宗教信念程度来理解这两道问题的答案。这两道问题帮助我们把人们分为三种(或者在宗教信念程度这一未观察到的变量上对他们赋量表值):

0:没有(或很少)宗教信念的人(不相信天堂或者地狱的存在)。

1:有一定宗教信念的人(相信天堂的存在但不相信地狱的存在)。

2:具有强烈宗教信念的人(同时相信天堂和地狱的存在)。

第五组问题涉及富裕程度。许多人拥有CD播放机,然而,至少在1997年荷兰的北部地区,只有一小部分人拥有洗

碗机(Sanders & van Schuur, 1998)。无论如何，拥有洗碗机的人一般也拥有CD播放机。因此这两道问题把人们分成了三类(或者在富裕程度这一未观察到的变量上对他们赋量表值)：

0：不是非常富裕(甚至没有CD播放机)。

1：富裕程度一般(只有CD播放机)。

2：更加富裕的(同时拥有CD播放机和洗碗机)。

第2节｜使用问题的答案作为测量工具

之前提到的五组问题可以作为工具来进行比较粗略的测量(即将人群分为三组)。测量值是定序的,意味着我们知道所测量的变量的值从一个组到另一个组逐渐增加,但是我们不清楚增加了多少。

测量值,或者量表值,可以简单地把每道题回答“是”的数量相加,来给每个人赋值。这样我们就得到了三个不同的分数0、1、2,可以用来进行分组。对于一个定序的量表而言,这种分类和用1、2、3进行分组是一样的,不过后者第一眼看起来更加合理。第一种分类(0, 1, 2)只是为了方便的原因。最小值永远是0,最大值和问题数目相等。数值“1”意味着,对于较为简单的问题A,回答者给出了正向的回答,而比较难的问题B则情况相反。稍后我们将处理不符合这一模式的情况(即回答者给出了问题B正向的回答,而问题A则相反)。

第 3 节 | 多于两个问题的定序测量

对于以上五个例子的每一个,我们都可以再加上无限的问题来测量同一个潜变量。举例来说,身高的问题可以用其他数字来测量,如下所示:

问题 C:你的身高超过 1.75 米了吗?	是/否

加上这个问题使我们可以把身高介于 1.70 米和 1.80 米的人进一步分成两组:身高处于 1.70 米和 1.75 米的组,及身高处于 1.75 米和 1.80 米的组。加上这第三个问题就把所有的回答者分成了四组。

每一组的量表数值——0、1、2、3——都叫做测量值。固然,它们不是用尺子测量得到的值,但是根据测量的统一定义:"给对象赋值,使得对象之间的关系就可以由数字之间的关系来代表。"(Coombs, Dawes, & Tversky, 1970:12)它们可以被视为测量值。在这个例子中,回答者之间的关系是他们相对身高之间的关系(一些人比另一些人要高)。并且数字之间的关系是简单的定序关系:数字从 0 到 3 排序,因此是定序量表。对回答者和问题进行排序是一种测量,没有用尺子进行测量的精准。在第 3 章,我们会具体阐述这一理念的结果。

类似地，计算能力可以通过增加更多的问题来更好地测量，比如下面这些：

问题 C:学生答对(237×21)/28 等于多少了吗?	是/否
问题 B:学生答对(13.267×108.936)/2.67 等于多少了吗?	是/否

如果我们假设 A、B、C、D 四个问题是根据难度的逐渐增加而排列的，那么一个人可以答对问题 D，他/她也可以答对问题 C、B 和 A。

对于这四个问题，我们现在做一个小数据集：编码“1”即“答案正确”，编码“0”即“答案错误”。如果问题是从简单到困难排序的，回答者是从能力不太强到能力较强来排序的，那么数据矩阵的下三角是正确的回答，上三角是错误的回答。这样一种结构被哥特曼称为量表图示法。

题项→	A B C D
回答者↓	
0	0 0 0 0
1	1 0 0 0
2	1 1 0 0
3	1 1 1 0
4	1 1 1 1

政治意识形态的问题(第三个例子)可以增加其他询问平等的问题，比如问题 C：做同样工作的男性和女性应该获得同样的收入吗？(是/否)。

假设(我们在实证研究中也发现这一点)有更多的人同意问题 C 而不是问题 A(也适用于问题 B)。用计算能力来做类比，我们说问题 C 比问题 A 和 B 更容易或者难度更低。当

问题与计算能力无关时,这一术语有时候比较容易使人迷惑。在这种情境下,容易和困难的意思只是指出有多少人给出了正向的回答(在我们的例子中是多少人回答了"是")。

为了扩展有关宗教信念的问题,我们可以增加更多的"你相信……吗?"的问题(或者题项,正如这些问题通常的称法)。比如,[2]

问题 C:你相信上帝吗?	是/否
问题 D:你相信死后重生吗?	是/否

为了扩展"富裕程度量表",我们可以使用询问其他财产的问题,例如,[3]

问题 C:你有彩色打印机吗?	是/否
问题 D:你有微波炉吗?	是/否
问题 E:你有车吗?	是/否

这里讨论的量表经常被称为累计量表。[4]个体的量表值越高,意味着在潜变量上的积累也越多。

第 4 节　哥特曼量表:理想的确定性累计量表

如果人们对一个难度高的二分题项给出正向回答的同时,也对所有较容易的二分问题给出正向回答,就没有太大问题。我们把这种情况下的一系列题项称为“确定性累计量表”。一个确定性量表不包含与量表相反的任何回答模式,因此有时也被称为理想量表。

让我们现在用符号和图示来正式介绍确定性量表模型。我们将展示:如果两个二分题项 **i** 和 **j**,分别代表潜在特征的连续统上具有不同基准或者门槛值,这意味着什么。如果我们把这个连续统理解为一个新变量,即,一个新的理论概念,那我们可以用一条水平线来代表它。这些二分题项用它们在这条线上的位置来代表;这些位置就表示了它们的基准。题项的位置用希腊字母 δ(delta)来表示,题项 **i** 用 δ_i 表示,题项 **j** 用 δ_j 表示。每个人在这一潜在连续统上都有自己的位置。我们用希腊字母 θ(theta)来表示个人,或者称为研究对象。“一般的”研究对象(研究对象 **s**)用 θ_s 表示。研究对象 **s** 对于题项 **i** 的回答把 **s** 定位在潜在连续统的一侧:回答“0”把 **s** 定位在 δ_i 的左侧,“1”把 **s** 定位在 δ_i 的右侧。如图 2.1 所示。



图 2.1 题项 i 把潜在连续统划分成两部分

如果有另外一个二分题项 **j**,相应的量表值是 δ_j,研究对象 **s** 在连续统上就具有了第二个位置,如图 2.2 所示。

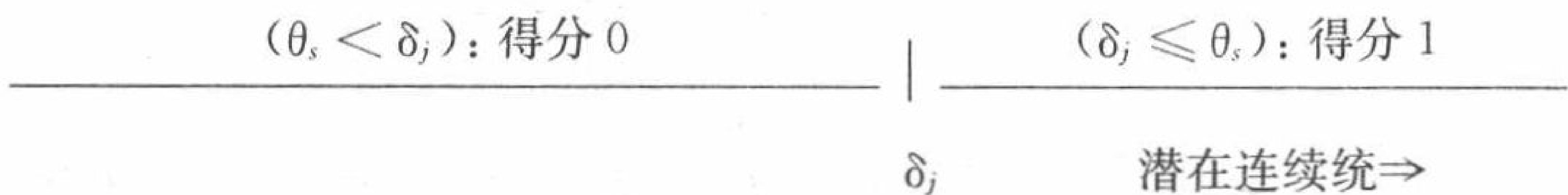


图 2.2 题项 j 把潜在连续统划分成两部分

如果我们同时在连续统上显示两个问题,那么就得到图 2.3。

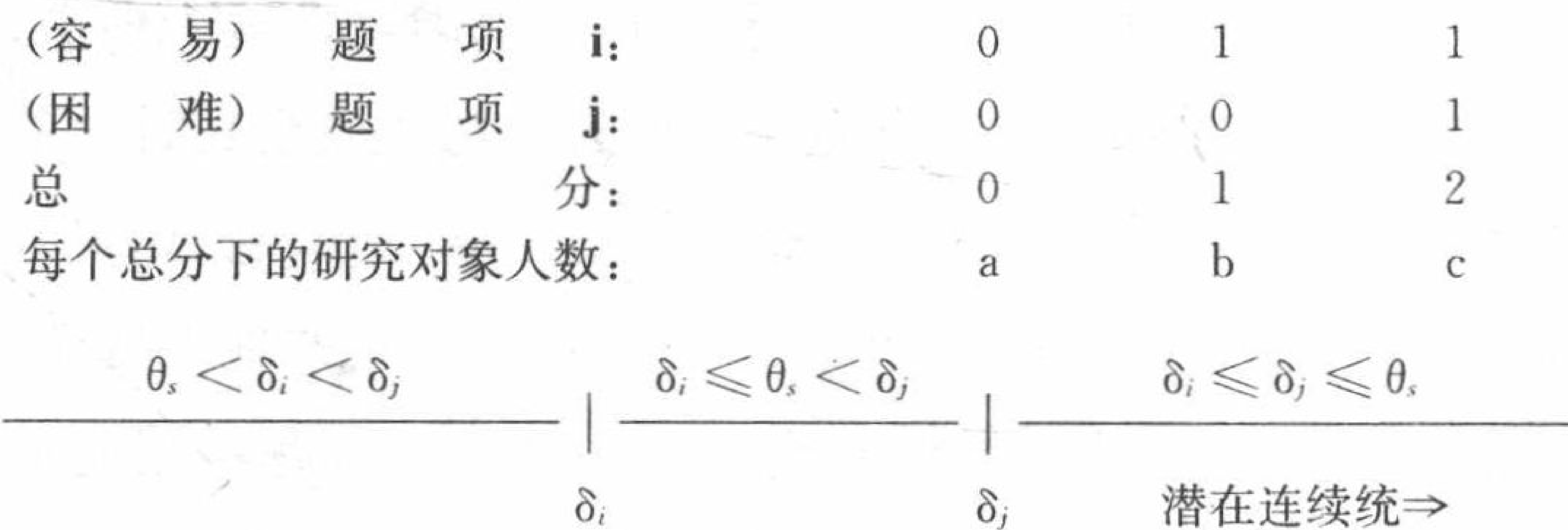


图 2.3 题项 i 和 j 把潜在连续统划分成三部分

图 2.1、图 2.2 和图 2.3 的信息可以用一个列联表来表示,题项 **i** 代表行,题项 **j** 代表列(参见表 2.1)。总分为 0 的研究对象人数(a)显示在(0, 0)单元格中,得分为 1 的研究对象人数(b)显示在(1, 0)单元格中,得分为 2 的研究对象人数(c)显示在(1, 1)单元格中。不存在题项 **i** 得分为 0 而题项 **j** 得分为 1 的研究对象。在理想的确定性累计量表中,(**ij**, 01)

的单元格是空白的。空白的单元格叫作误差单元格。每个研究对象的量表数值等于他/她给出正向回答(1)的问题总数。

表 2.1 两个理想累计二分题项的列联表;单元格包含了对题项 i 和题项 j 的回答频数

	题项 j			
		0	1	总分
题项 i	0	a	0	a
	1	b	c	b+c
总分		a+b	c	a+b+c

第5节 | 确定性模型的假设

到目前为止,我们介绍的确定性模型具有一系列假设。

首先,要测量的潜在特征是一个单一的特征,并且可以由一个单一维度的连续统来表示。

第二,如果一个研究对象的量表值比题项的量表值低,那么该研究对象对这个问题做出了负向回答;如果一个研究对象的量表数值不低于或者比题项的量表值高,那么该研究对象对这个问题做出了正向回答。

第三,对每个问题做出正向回答的概率只取决于研究对象在潜在特征上的取值,不受其他任何系统性的影响(即,局部随机独立性假设)。

第6节 | 路易斯·哥特曼和哥特曼量表

长期以来占统治地位的模型是累计量表分析，蕴含量表分析，以及哥特曼量表法。它是由社会学家路易斯·哥特曼在第二次世界大战期间发展出来的(关于定量社会科学研究丛书系列中哥特曼量表的更多信息，请参阅 Andrich，1988；Jacoby，1991；或者 McIver & Carmines，1981)。作为一位社会科学家，哥特曼被要求评价美国军人的素质和士气(Guttman，1950)。他设计了一系列问题，询问士兵在战斗中的恐惧，对于长官的态度，以及其他此类问题。[5]哥特曼的“对战斗的恐惧”量表包括下列问题：

当你处在战斗状态下，你具有下列回答的频率如何？

A. 胃部难受　　经常/不经常

B. 呕吐　　经常/不经常

C. 小便失禁　　经常/不经常(“经常”是正向选项)

哥特曼的“对长官的态度”量表包含以下问题：

A. 与士兵相比，你觉得长官们的特权：

1.长官具有太多的特权；2.长官具有稍微多的特权；3.长官们的特权适当；4.长官们的特权太少。(正向

选项:1 和 2)

B. 你认为长官认识到你的能力和潜能了吗?

1.是的,我很肯定;2.是的,我想是的,但是不肯定;3.没有,我想他们没有认识到;4.不清楚。(正向选项:3)

C. 你认为有多少长官滥用了他们的军衔级别?

1.几乎所有;2.大多数;3.一些;4.一少部分;5.没有。(正向选项:1 和 2)

遗憾的是,在社会科学研究中理想的确定性哥特曼量表很难存在。所以接下来我们转到非理想的量表。

第3章

非理想的累计量表

第1节 | 模型违反

如果一系列题项没有构成一个理想的哥特曼量表,而是包括了一些“错误”的回答,我们也没有必要舍弃它。“错误”的回答,或者说“模型违反”“模型误差”,或者简单的“误差”,是指与模型的含义不一致的回答。哥特曼的量表模型非常具有限制性。k 个二分题项总体上具有 2^k 个可能的回答模式,但是只有 $k+1$ 个回答模式构成了一个理想的哥特曼量表。所以对于8个二分题项,存在 $2^8=256$ 个可能的回答模式,但是只有9个可接受的模式。即使对于最好的一组问题,数据集只包括这9类可接受的回答模式是非常不可能的。因此我们需要考虑如何定义模型违反以及可以接受的模型违反的数目是多少。

关于如何定义超过两个题项的回答模式的误差或模型违反的数量,在20世纪五六十年代引起了广泛的讨论。比如,对于A、B、C、D四个题项的回答模式ABCD, 1101,其中A、B、C、D按照从易到难的顺序排列。如果正向回答的数目是一定的,那么这里是存在两个误差吗?因为题项C应该是正向回答而题项D应该是负向回答?或者这里只有一个误差,因为改变C或D的任何一个回答都会得到更完善的回答模式?

更加让人困惑的是，我们还要问为了测量的目的是否能够继续使用 ABCD，1101 这一回答模式？如果是，该如何使用？以这种方式回答的研究对象是否应该得到量表值为 2？因为通过把题项 D 的回答变为 0，他/她的回答模式就更完善。又或者他/她是否该得到量表值为 4 呢？因为把题项 C 的回答更改为 1 同样也得到更加完善的回答模式。或者他/她是否该得到量表值为 3 呢？因为他/她对三个问题给出了正向回答。莫坎在其文章第 2 章中对此种争论做出了出色的回顾（Mokken，1971）。我们将在本章稍后部分讨论这个问题。

第 2 节 | 误差:违反题项和研究对象之间的传递关系

个人对一个问题的回答被认为是个人和该问题(题项)间的一种支配关系。如果他/她给出了正向回答,则个人支配题项,反之则是题项支配个人。如果我们有更多的研究对象和更多的题项,并且所有研究对象和题项可以沿着一个单一的维度分布,我们就可以推论出其他两种类型的支配关系,即研究对象之间的支配关系和题项之间的支配关系。之所以可以这样表达,是因为当所有的研究对象和题项能够在一个单一维度上表示的时候,所有的支配关系都具有传递性。在这种情况下,我们可以在这一维度上同时对研究对象和题项进行排序。通过展示这一传递关系如何被违反,我们可以更好地理解什么叫做模型违反,或者误差。

让我们假设有三个题项,A、B 和 C,按照难度排列,以及四个研究对象,W、X、Y 和 Z,按照能力排列(参见图 3.1):

图 3.1 四个研究对象(W, X, Y, Z)和三个题项(A, B, C)在同一累计量表上的分布

1. 我们可以用下列方式来指定三个题项之间的传递关系：如果题项 B 比题项 A 难度高，并且题项 C 比题项 B 难度高，那么 C 比 A 难度高。这一传递关系在逻辑上是正确的，如果我们用给出正向回答的研究对象的频次（或者比例，作为相对频次）来定义“容易”和“困难”。如果两个或者更多的题项具有相同比例的正向回答，我们无法进行排序，它们是连在一起的。拿表 2.1 来讲，单元格（0, 1）和（1, 0）应该是空白的。但是如果人们同意每个题项的比例是不同的，那么我们可以根据“普及度”或者相反地根据“难度”，对题项进行排序。

现在我们可以区分其他三种传递关系。

2. 两个研究对象和一个题项之间的传递关系。如果 Y 的能力高于题项 B 的难度（即，Y 在单一维度上的位置处于题项 B 的右侧），并且 X 的能力低于题项 B 的难度（即，X 在单一维度上的位置处于题项 B 的左侧），那么 Y 的能力高于 X 的能力。这种传递能力（Y > B, B > X，因此 Y > X）被用来对两个人的能力或者在某一维度上的量表值进行排序。这一传递关系的结果可以在第三种传递关系中应用。

3. 三个研究对象间的传递关系。如果 Y 的能力强于 X，并且 Z 的能力强于 Y，那么 Z 的能力强于 X。这种传递关系使得我们可以用排序作为测量值对所有人进行赋值（即，用在一个定序量表上的量表值对所有人进行赋值）。

第四种传递关系非常重要。

4. 一个研究对象和两个题项之间的传递关系。如果 Y 的能力高于题项 B 的难度，并且题项 B 的难度高于题项 A 的难度，那么 Y 的能力也高于题项 A 的难度。换句话说，如果 Y 能够对较难的题项 B 给出正向回答，那么他应该也能够对

较容易的题项 A 给出正向回答。注意我特意使用“应该”一词,因为上述情况并不总是真实的。

在以上四种传递关系中,只有最后一种可以在逻辑上被违反。因此我们将模型违反定义为对于个人和多个题项之间的传递关系的违反。个人对较难题项给出了正向回答时,却对容易题项做出了负向回答,则他/她违反了累计模型;也就是说,他/她的回答在确定性累计模型上存在误差。这就是为什么这种误差有时候被称为“哥特曼误差”。个人在回答模式 ABCD, 1101 上所犯的误差数目为 1,因为只有题项对 CD 违反了累计模型。[6]

下面让我们来看如何用下列题项来计算模型违反的数量(题项按照难度从左到右来排序)。四个题项具有六个成对的题项组合,4×(4−1)/2。题项违反的数量是违反模型的题项对的数量。表 3.1 给出了一些例子。题项的次序是按

表 3.1 按照难度从易到难排列的四个题项中,违反累计模型的题项对(1:违反;0:没有违反)

ABCD	*AB*	*AC*	*AD*	*BC*	*BD*	*CD*	
0011	0	1	1	1	1	0	4 个误差:题项对 AC, AD, BC, BD
0111	1	1	1	0	0	0	3 个误差:题项对 AB, AC, AD
0101	1	0	1	0	0	1	3 个误差:题项对 AB, AD, CD
0001	0	0	1	0	1	1	3 个误差:题项对 AD, BD, CD
1011	0	0	0	1	1	0	2 个误差:题项对 BC, BD
0110	1	1	0	0	0	0	2 个误差:题项对 AB, AC
1001	0	0	0	0	1	1	2 个误差:题项对 BD, CD
0010	0	1	0	1	0	0	2 个误差:题项对 AC, BC
1101	0	0	0	0	0	1	1 个误差:题项对 CD
1010	0	0	0	1	0	0	1 个误差:题项对 BC
0100	1	0	0	0	0	0	1 个误差:题项对 AB
违反数量	4	4	4	4	4	4	在该(假设的)数据集中存在 24 个模型违反

照难度事先排列好的，A 是最容易的，D 是最难的。在第一个回答模式中（ABCD，0011），六个题项对中的四个违反了模型。这一回答模式的模型违反，或者研究对象所犯的模型违反，则为四个。

第3节 | 在较大数据集中误差定义的扩展

当研究对象回答模式所犯的误差数量被定义为其回答模式中违反累计模型的题项对的数量时,确定数据集中误差的总数量就很变得很容易:所有研究对象所犯误差的数量总和。如表3.1所示,如果数据集包含了11种回答模式,那么该数据集中一共存在24个误差,这个数字可以从最后一列或者最后一行计算得到。

我们也可以分别计算每一个题项中存在的误差数。在设计一个累计量表的后期或者评价一个累计量表中的所有题项是否都一样好的时候,知道每一个题项的误差数是非常有用的。虽然我们需要成对的题项来定义误差,但是我们可以通过把误差分别归因到每一个具体的题项来区分成对的题项的数目。每一个题项的误差数量被定义为加总了所有人的包括该题项和一个误差的成对题项的数量。

表3.2给出了一个小例子。表3.1比较了每一个题项对,表3.2则是通过合并包括每个题项的题项对来考察单个的题项。题项A涉及题项对AB, AC和AD,题项B涉及题项对AB, BC和BD,以此类推。对于回答模式是0001的第一位研究对象来说,三个题项对违反了模型:AD, BD和CD。

因此题项 A，B 和 C 每一个都被涉及了一次，而题项 D 则被涉及了三次。这里我们只展示了五个回答模式，而没有展示具有一个或多个误差的由 4 个题项组成的 11 个回答模式。

表 3.2　分别对每个题项计算误差总数

题项中的误差(即，包含该题项的题项对)							
	ABCD	*A*	*B*	*C*	*D*	总数	题项对
研究对象 1	0 0 0 1	1	1	1	3	3	AD，BD，CD
研究对象 2	0 1 1 1	3	1	1	1	3	AB，AC，AD
研究对象 3	1 0 0 1	0	1	1	2	2	BD，CD
研究对象 4	1 0 1 1	0	2	1	1	2	BC，BD
研究对象 5	0 0 1 0	1	1	2	0	2	AC，BC
总　　数		5	6	6	7	12	

第4节 如何评价数据集中的模型违反数量

既然我们已经定义了模型违反或者模型误差,也计算了这些误差的数量,那么我们如何评价误差的数量呢?“没有太多误差”是什么意思?文献已经给出了很多不同的回答,并且莫坎在书里提供了一个出色的综述(Mokken, 1971)。如果能够把观察到的误差数量与一个基准进行对比,我们就可能用模型拟合这一测量来表示模型误差的数量。模型拟合的若干种测量,也可以叫做可量测性的标准,已经被发展出来了。我们现在就讨论几种主要的测量。

我们用来比较数据集中误差数量的基准是什么?一个直观的答案是数据集中可能存在的误差的总数量。但是人们能够犯的最大数量的误差是多少呢?

最坏的情况是每个人对所有高难度的题项都给出正向回答,同时给所有简单的题项给出负向回答,比如 ABCD, 0011。但是在几乎所有的实际情况中,题项的难度是由所研究的数据集决定的。因此如果每一个人都给出了 ABCD, 0011 的回答,那么题项 C 和 D 就被自动定义为容易的题项,而 A 和 B 则被定义难度高的题项。在这类情况下,误差的最大数目永远不可能等于回答的总数目,即人数乘以题项数

$(N\times k)$。事实上,定义数据集中可能存在的最大误差数目不是那么容易。很多程序可以给我们提供该数目的估计值,但是它们都被证明高估了真实的最大误差数量。

如果我们能够决定可能的最大误差数目,那我们可以通过比较观测到的误差数目和最大误差数目来发展一个可量测性标准。由哥特曼发展的第一个可量测性标准叫做再现性系数,**Rep**。

$$\mathbf{Rep}=1-\frac{\mathrm{Err(obs)}}{N\times k} \qquad [3.1]$$

在这里,Err(obs)是观察到的误差数量,$N\times k$ 是回答的总数量。**Rep** 可以被理解为没有误差的回答的比例。

在之后发展出的可量测性标准里,$N\times k$ 被一个更好的估计所替换,即模型违反最大数量 Err(max)。可量测性系数 **S** 于是被定义为:

$$\mathbf{S}=1-\frac{\mathrm{Err(obs)}}{\mathrm{Err(max)}} \qquad [3.2]$$

但是即使我们有了一个对最大误差数目的合适估计值,也有其他的原因使得我们不想比较观测到的误差数量和最大误差数量。为了很好地理解这一点,我们可以把建立可量测性指标想象为一种假设检验。通常我们通过比较一个假设与其他假设来进行假设检验。一个假设是我们的数据集确实构成了一个累计量表。但是其他的假设应该是什么呢?当我们比较观察到的误差数目和可能的最大误差数目时,在一种可能的题项排列顺序下,其他的假设在根本上可以被理解为:是数据符合一种与累计量表尽可能不同的模型。但是,这种“其他假设”究竟有何用呢?

一个更加合适的其他假设是零假设,它假设所有的题项毫无关系。它们并不构成一个累计量表,也不构成其他某些奇怪的极端模型。如果题项是彼此无关的,我们就不能预测对较难题项给出正向回答的研究对象会对较容易题项给出正向回答。换句话说,我们关于研究对象是否对一条容易的题项做出正向回答的预测不会随着我们是否知道该研究对象对较难题项给出正面回答而改变。

洛文杰(Loevinger)在1948年建议应该比较观测到的误差数目和在统计独立性条件下的期望误差数量,Err(exp),但是直到20世纪60年代末才被莫坎所采用。莫坎重新介绍了洛文杰提出的同质性指数**H**作为一条可量测性的标准:

$$\mathbf{H}=1-\frac{\mathrm{Err(obs)}}{\mathrm{Err(exp)}} \tag{3.3}$$

在这里对于题项对(**i**, **j**)而言,Err(exp)是在统计独立性条件下的误差的期望数量,Err(obs)是同时对难度较高的题项**j**给出正向回答但对较容易题项**i**给出负向回答的研究对象数量[7]。**H**=1时模型拟合最好,因为不存在任何误差。**Rep**=1或者**S**=1也代表了最好的模型拟合。**H**=0则意味着这时我们的数据集是一个完全随机的数据集。当我们在统计独立性条件下所观察到的误差数量多于预期值时,**H**可以是负值。这种情况是可能发生的,比如当我们对题项的难度进行了错误的排序。

现在我们来比较下面两个假设:

零假设	H_0	题项之间是不相关的。
模型假设	H_1	题项构成了一个累计量表。

在统计学中拒绝或者证伪零假设而支持模型假设是非常普遍的做法。模型假设通常被叫做备择假设。可以这么来理解模型假设检验,即比理想状况糟糕的所有假设都是不好的,但是比随机状况要好的所有假设都是好的。

在统计独立性条件下描述误差数量的期望值最容易的方式,就是做两个题项的交互表。举例来说,我们回到之前关于宗教信念的两个问题上去:

问题	回答
问题 A:你相信天堂吗?	是/否
问题 B:你相信地狱吗?	是/否

让我们假设访谈了 100 个人,24 个人相信天堂和地狱,34 个人既不相信天堂也不相信地狱,36 个人只相信天堂但不相信地狱,还有 6 个人相信地狱但是不相信天堂(参见表 3.3a)。这两个题项构成了一个累计量表吗?让我们来比较在统计独立性条件下观察到的误差数量与误差数量的期望值。

表 3.3a 假设的实际情境:对天堂的信念 vs.对地狱的信念

		天堂		
		是	否	总数
地狱	是	24	6	30
	否	36	34	70
	总数	60	40	100

我们怎么确定观察到的误差数量呢?在这个 2×2 的交互表中,行变量是难度高的题项(只有 30 个正向回答),列变量是难度低的题项(60 个正向回答),右上角的单元格(对地狱答"是",对天堂答"否")可以被看做是误差单元格。在一

个理想的累计量表里,这个单元格应该是空的,像第二个交互表展示的那样(参见表3.3b),这是一种理想情况。在实际情况下,这个误差单元格包括了6个人,即,Err(obs)=6。

表3.3b 适用于表3.3a中的题项的理想的哥特曼量表

		天堂		
		是	否	总数
地狱	是	30	0	30
	否	30	40	70
	总数	60	40	100

我们怎么找到在统计独立性条件下的误差数量的期望值?在统计独立性条件下,对于两个题项的既定回答的概率等于每条题项的回答概率的乘积。因此如果答对难度高的题项的概率是0.30,答错较容易题项的概率是0.40(1−0.60),那么这种回答模式的概率是0.3×0.4=0.12。因为有100个人,所以我们期望有100×0.12=12个人会对这两个题项给出错误回答,如果题项之间在统计上是独立的。因此Err(exp)=12(参见表3.3c)。

表3.3c 统计独立性条件下表3.3a题项回答的交互表

		天堂		
		是	否	总数
地狱	是	18	12	30
	否	42	28	70
	总数	60	40	100

现在我们计算这两个题项组成的量表的同质性系数:

$$\mathbf{H}=1-\frac{6}{12}=0.50$$

如果分母 Err(exp)为 0,我们不可能计算 **H** 的值。当所有的研究对象都对较容易的题项给出正向回答或者对较困难的题项给出负向回答时,Err(exp)的值为 0。每一个研究对象都给出相同答案的题项不被包括在分析里。

同质性系数为 0.50。现在下一个问题是:“这个值是高还是低?”

第 5 节 | 评价同质性系数

有两种方式来回答这一题项,"**H** =0.50 的同质性系数是高还是低?"通俗地,我们可能会问:"0.50 和 0.00 差别有多大?"或者完全没有同质性;或者我们还会问,"0.50 和 1.00 差别多大?"或者完全的同质性。对于第一个问题,有统计意义上的答案。在同质性系数为 0.00,即题项回答之间毫无关系的总体中,有可能找到一个具有特定同质性系数(比如 0.50)的规模为 N 的样本,并且我们可以估计出这一概率。具体可以参见附录第 1 部分。这种方式要在统计独立性条件下,找到 **H** 系数的分布,并且推算出一个(单边)介于 0 和一个正数之间的置信区间,特定的超越概率设定为 α(通常为 5%)。如果 **H** 落在这一置信区间内,我们就接受零假设,拒绝模型假设。这种决定可以用 Z 和一个 $Z(i)$ 的统计量来决定,Z 适用于整个量表,$Z(i)$ 适用于题项 i:如果 Z[或者 $Z(i)$]的值足够大(一般 > 3),那么整个量表的同质性(或者题项 i 的同质性)就不能仅仅用偶然性来解释。

但是接受模型假设——总体中的同质性系数高于 0.00——并没有就第二个问题给出答案:多高的值算是高?或者我们的数据集离理想的同质性有多远?遗憾的是,对于这一问题并没有简单的答案。统计意义上显著的相关性或

者同质性系数可能不是非常重要的。莫坎提出系数低于0.30的数据集的同质性不够高,不能构成一个累计量表。他的这一建议建立在大量的经验以及与基于信度和因子分析的量表的比较基础之上。

第6节 在具有超过两个题项的量表中使用同质性系数

既然我们可以用同质性系数来测试两个题项是否可以构成累计性量表,我们可以将此测试扩展到包括大于两个题项的量表。在多个题项构成的量表中,我们把每个题项对里观察到的误差数目 Err(obs)加总。因此对四个题项而言,需要加总六个题项对的误差数目。我们也可以在统计独立性条件下为每一个题项对计算期望误差数目 Err(exp),或者进行加总。整个量表的同质性系数,**H**,用下面的公式来表示:

$$\mathbf{H}=1-\frac{\sum_{i=1}^{k-1}\sum_{j=i+1}^{k}\mathrm{Err(obs)}}{\sum_{i=1}^{k-1}\sum_{j=i+1}^{k}\mathrm{Err(exp)}} \qquad [3.4]$$

下面用四个题项作例子,如表3.4和表3.5所示:

所有题项对观察到的误差数目是5+6+4+10+7+8=40。

所有题项对的期望误差总数是15+12+6+20+10+12=75。

$$\mathbf{H}=1-\frac{40}{75}=0.47$$

表 3.4　四个题项的例子

		题项 B					题项 C					题项 D		
		是	否				是	否				是	否	
题项 A	是	25	5	30	题项 A	是	24	6	30	题项 A	是	26	4	30
	否	25	45	70		否	36	34	70		否	54	16	70
		50	50	100			60	40	100			80	20	100
		题项 C					题项 D					题项 D		
		是	否				是	否				是	否	
题项 B	是	40	10	50	题项 B	是	43	7	50	题项 C	是	52	8	60
	否	20	30	50		否	37	13	50		否	28	12	40
		60	40	100			80	20	100			80	20	100

表 3.5 对于表 3.4 的结果汇总

题项对	*AB*	*AC*	*AD*	*BC*	*BD*	*CD*	总和
Err(obs)	5	6	4	10	7	8	40
Err(exp)	15	12	6	20	10	12	75
H	0.67	0.50	0.33	0.50	0.30	0.33	0.47

现在也可以对每个题项,A、B、C 和 D,计算可量测性系数。在这种情况下,观测的和期望的误差数量的计算需要加总所有包含了所有要计算系数的题项的题项对。

$$\mathbf{H}_A = 1 - \frac{5+6+4}{15+12+6} = 1 - \frac{15}{33} = 0.55$$

$$\mathbf{H}_B = 1 - \frac{5+10+7}{15+20+10} = \frac{22}{45} = 0.51$$

$$\mathbf{H}_C = 1 - \frac{6+10+8}{15+20+12} = 1 - \frac{24}{47} = 0.49$$

$$\mathbf{H}_D = 1 - \frac{4+7+8}{6+10+12} = \frac{19}{28} = 0.32$$

每个题项的同质性系数都高于最低界线 0.30。很容易证明,当一个累计性量表中的每一个题项的同质性系数都高于某数值 c,那么该量表整体的同质性系数也高于某数值 c。对单个题项同质性系数的考察可以让研究者评估某个题项是否适合纳入累计性量表中。同质性不够强的题项不应该成为累计性量表的一部分。举例来说,题项 D,刚刚超过最低数值 0.30。表 3.6 给出了另外一个计算的例子。

表 3.6　累计性量表的一个计算小例子

累计性量表的误差数量										
A	*B*	*C*	*D*	频数	*AB*	*AC*	*AD*	*BC*	*BD*	*CD*
1	1	1	1	70						
1	1	1	0	240						
1	1	0	1	40						40
1	0	1	1	20				20	20	
0	1	1	1	8	8	8	8			
1	1	0	0	160						
1	0	1	0	60				60		
1	0	0	1	28					28	28
0	1	1	0	16	16	16				
0	1	0	1	14	14		14			14
0	0	1	1	4		4	4	4	4	
1	0	0	0	168						
0	1	0	0	48	48					
0	0	1	0	24		24		24		
0	0	0	1	10			10		10	10
0	0	0	0	90						
—	—	—	—	—	—	—	—	—	—	—
726	596	442	194	1 000	86	52	36	108	62	92
期望误差数目					163	121	53	179	115	108
H(ij)：					0.47	0.57	0.32	0.40	0.46	0.15

续表

题项对 AB 的累计性量表存在的期望误差数量为 (1 000−726)×0.596 = 163.30。
H(AB)的计算方法为 1 − 86/163 = 0.47。
累计性量表的题项系数：

E(o)A：86 + 52 + 36 = 174	E(e)A：163 + 121 + 53 = 337
H(A) = 1 − 174/337 = 0.48	
E(o)B：86 + 108 + 62 = 256	E(e)B：163 + 179 + 115 = 457
H(B) = 1 − 256/457 = 0.44	
E(o)C：52 + 108 + 92 = 252	E(e)C：121 + 179 + 115 = 408
H(C) = 1 − 252/408 = 0.38	
E(o)D：36 + 62 + 92 = 190	E(e)D：53 + 115 + 108 = 276
H(D) = 1 − 190/276 = 0.31	
Total Err(obs) = 872	Total Err(exp) = 1 478

$\mathbf{H}_{(scale)}$ = 1 − 872/1 478 = 0.41
对于 total Err(obs)和 total Err(exp)，我们可以除以 2，因为每个题项对被计算了两次。

对于大于两个题项的量表，我们不仅可以计算是否所有的 **H**(ij)显著地大于 0，也可以计算是否所有的 **H**(i)和总体的 **H** 是否显著地大于 0(参阅附录第 1 部分)。

使用 0.30 作为每个题项和量表同质性系数的最低线通常是远远高于统计显著性的边界的。0.30 的同质性系数不显著的情况只有在被访者数量非常少(比如低于 50 个)，题项的数量非常少(比如 2 个或 3 个)，以及题项的难度是极端值(比如大于 0.90 或者低于 0.10)才会出现。当不存在这些情况时，一般没有必要检验零假设。我们之所以要把同质性系数的最低线定得高于统计显著性的边界，是为了易于理解或者有实质性的关联。依据我们的经验，同质性系数低于 0.30的量表或者题项很难去解释。

第7节 误差的“原因”:题项还是研究对象?

当研究对象对题项的一组回答不符合理想的哥特曼量表时,目前为止我们采用了“归咎于”(blame)题项的方法:题项不是对潜在特征的足够好的测量指标。但是在哥特曼模型中,一个误差仅仅是简单地违反了对于研究对象和题项对之间的期望关系。因此所谓的误差既可能是由于研究对象采用了一套不同的行为准则,抑或是不同的研究对象对同样的问题的理解有差异。

如果我们希望一个测量工具可以适用于不同的研究对象,不同的时间段,或者不同的实验场景,那我们最好找到一个量表,其中所有的题项对于所有的研究对象的作用机制都相同。如果有必要,我们可能要放弃那些作用机制不同的题项,只保留最大数目的还有用的题项。剩下的这些题项仍然可以被视为对于潜在特征变量的典型测量指标。

我很难有足够的理由去忽视一些研究对象,而仅仅专注于大部分的研究对象。这样的做法带来的问题是:研究者是从一个不完整的样本推论更大的总体。原来的样本通常是从一个使用严谨方法来定义的总体中抽出来的。如果我们

现在删除一些不适合量表的研究对象,剩下的样本可能不再具有很好的代表性。因此我们通常更倾向于删除题项而不是研究对象。当然,在某些情形下可能需要辨别并删除不正常的样本。下面我们介绍一个此类方法。

第8节｜“归咎于”研究对象：转置数据矩阵及计算研究对象同质性

确定哪些研究对象做了最多的模型违反的一种方法就是，简单地计算在每一种回答模式里计算违反哥特曼量表的误差数量。但另一个方法利用了在一个理想的哥特曼量表中，题项和研究对象的角色是完全对称的，因此他们可以互换的优势。所以，我们可以简单地转置数据矩阵，即互换行和列，如表3.7a和表3.7b所示。在一个理想的哥特曼量表中(表3.7a)，无论我们是采用通常的做法将研究对象作为行、题项作为列，还是把研究对象作为列、题项作为行，我们都能得到理想的回答模式。在两种情况下我们都看到了下三角矩阵全为1，上三角矩阵全为0。

题项和研究对象之间的对称强调了我们也可以通过研究对象的个体同质性来评价他们。与计算同质性系数的方法相同，我们可以对整个数据集计算一个同质性系数 $\mathbf{H}^{\mathrm{T}}$，对个体研究对象计算 $\mathbf{H}^{\mathrm{T}}(s)$，以及计算一对研究对象 s 和研究对象 t 的同质性系数 $\mathbf{H}^{\mathrm{T}}(st)$。主要的区别在于，通常计算题项同质性是针对上百个研究对象进行的，而计算研究对象同质性只是针对5到20个题项。因此对于这些研究对象同质性的估计值应该持保留态度。

梅杰已经展示了哥特曼量表的误差数目是一个简单而有力的适用于人的统计量,并且与其他的统计值比较起来性能也比较好,如研究对象同质性(Meijer, 1994)。但计算研究对象同质性对于其他概率性模型的评价也很有用,我们在本书稍后部分会讨论到。

表 3.7a 理想的哥特曼量表,研究对象 1—6 为行,题项 A—E 为列

数据矩阵					
	A	B	C	D	E
1	0	0	0	0	0
2	1	0	0	0	0
3	1	1	0	0	0
4	1	1	1	0	0
5	1	1	1	1	0
6	1	1	1	1	1

表 3.7b 理想的哥特曼量表,题项 E—A 为行,研究对象 6—1 为列

转置的数据矩阵						
	6	5	4	3	2	1
E	1	0	0	0	0	0
D	1	1	0	0	0	0
C	1	1	1	0	0	0
B	1	1	1	1	0	0
A	1	1	1	1	1	0

第9节 | 使用非理想模式来测量研究对象量表值

我们可以使用具有误差的回答模式来对研究对象进行测量吗？答案可以为“是”，如果我们仍然可以假设所有的题项构成了一个累计性量表，并且研究对象和我们试图测量的其他人以同样的方式来理解问题。如果我们能够使用每个回答模式来测量研究对象，不考虑他所犯的误差数量，那么我们如何获得这一测量呢？

由于我们没有根据改善回答模式所需要改变的条目数目来定义一个回答模式中的误差数目，我们就不能把最邻近的理想回答模式的量表值赋给它。我们同时也看到了，比如对于回答模式 ABCD，1101 而言，最近的理想模式并不明确；可能为 2 或 4。因此我们倾向于把给出一个包括误差的回答模式的研究对象的量表值简单定义为给出正向回答的题项数量。[8]因此具有 ABCD，1101 回答模式的研究对象的量表值为 3。

第 10 节 | 结论

在本章中我们讨论了如何使用洛文杰的同质性系数来评价一个作为累计性量表的数据集。另外也讨论了如何测量研究对象和题项:用量表分值的排序来测量研究对象,用数据集中的普及度的排序来测量题项。下一章将讨论的问题是,在并不是全部的题项都能构成累计性量表的情况下,如何找到构成累计性量表的部分题项,以及在该种情形下是否应该删除部分题项或者研究对象。

第4章

确定或寻找构成量表的题项

第1节 | 寻找可以构成量表的题项

目前为止我们是把一组题项想当然地认为是一个累计性量表。有人可能说我们已经测试过这组题项是否可以构成量表,并且这种程序就是*测试或确认程序*。但是在实际情况中,我们并不会提前知道一套题项是否构成累计性量表。因此我们常常需要探索是否所有的或者只有一部分的题项符合累计性量表的要求。在本部分我会解释如何在许多题项中寻找累计性量表的最大可能性,即,能够包含最多题项的具有累计性量表(同时参考 Jacoby, 1991:35—37,供讨论)。这种方法叫做*寻找或者探索方法*。

该方法分为两步。第一步,我们找到最佳可能性的最小量表。我们可以根据理论、前人研究、文献或者直觉来定义最小量表。然而最经常的做法是,我们通过一种标准程序来选择最佳的最小量表。第二步,我们在既有量表中加入下一个最可能构成量表的题项。只要题项符合可量测性的标准,我们就不断地往已经存在的量表中一个一个地加入题项。

该方法可以视为分层级的综合(或者由下而上)聚集法。这种选择题项的程序与在其他模型中使用的自上而下的方法显然不同,后者如可靠性分析,因子分析,或者参数 IRT 模型,在这些的方法中,我们从整个量表、整个因子或者整个测

试中删除不好的题项。

最佳可能性的最小累计性量表包括两个题项。如果我们有两个题项,很可能就证伪这两个题项能构成一个量表。我们根据下列标准来定义两个题项的可量测性:

1. 如前所述,**H**(ij)系数应该显著地大于0,即$Z(ij) > 1.64$(单尾检验,置信度$\alpha = 0.05$),如前所述(参考附录第2部分)。这点意味着两个题项是正相关的。

2. **H**(ij)系数应该高于一个使用者自定义的最低界限,通常定为0.30。

这两个标准是有所不同的:第一个可以被视作一个统计显著性的标准,第二个则是实质相关性的标准。**H**系数可以显著大于0,特别是当样本足够大的时候,但是它的实际数值可能非常小。第二个标准则试图通过设定一个绝对的最低界限来纠正这一问题,比如0.30。

3. 在所有符合前两条标准的所有题项对中,选择具有最高的**H**(ij)系数的题项对。

4. 如果两个或者多个题项具有相同的最高的**H**(ij)系数,选择包括最少的普遍性题项的题项对。这条主观标准能够保证永远选择那些独特的题项对。

5. 如果还有很多题项对都符合这些标准,选择那些包含了在数据集中最先描述的题项的题项对。这最终保证我们选择到一个独特的最佳的最小规模的量表。

这个最小规模的量表,或者在此种意义上的任一量表,现在可以对其进行扩展,添加更多的题项。这个步骤是逐步进行的,每次只添加一个题项。需要添加的最佳题项需要满足下面情况:

1. 包括最好的下一个题项以及已经在量表中的某一条题项的所有题项对,都必须具有一个正向的 **H**(ij)系数。假如在所有可能的 **H**(ij)系数中,只有一个是稍微负向的,该题项也不会被选中以构成量表。这一要求说明量表中所有的题项必须是正相关的。

2. 新的题项的 **H**(i)系数必须显著地大于 0,即它的 $Z(i) > 1.64$,或者相关的使用者自定义的超越概率不是 0.05。

3. 新的题项的 **H**(i)系数必须大于使用者自定义的最低界限(通常是 0.30)。

4. 量表总的 **H** 系数必须大于使用者自定义的最低界限。

5. 如果不只一个题项满足前三条标准,选择能够给总的量表带来最高的 **H** 系数的题项。

6. 如果不只一个题项满足前四条标准,在备选的所有题项中选择具有最高的 **H**(i)系数的题项。

7. 如果还是有多过一个选择,选择难度最大的题项。

在筛选的过程中可能会出现三个问题。第一,随着越来

越多的题项被加入到量表中，原先所有题项的**H**(i)系数可能会下降，甚至低于使用者定义的最低界限。但是完成筛选分析的电脑程序不会(还不会)提示这个问题。一个可能的解决方式是，如果加入某一条题项可能会导致现有的**H**(i)系数下降幅度过大，那么最好不要加入该题项。但是我们并不推荐这一解决方法，因为它过分强调了之前存在于量表中的题项，而这种强调并没有充分的理由。因此，使用者必须检查在最终的量表中的每一条题项的**H**(i)系数值，以确定每一条题项依然具有足够的同质性。如果某一题项的**H**(i)值低于最低界限，可能就需要把它剔除掉。

第二个可能的问题则是纯粹的偶然性题项。某条题项看起来是一个很好的备选，可能纯粹出于偶然性，即，它碰巧和量表中既存的题项之间的**H**(ij)系数高于0。为了最大限度地避免这一问题，我们需要令加入一个新题项变得更加困难，以减少偶然性题项的发生。一种方式是当加入新的题项时，就降低原先由研究者规定的超越概率(比如0.05)，附录第2部分介绍了此种方法。

第三种可能存在问题是，前面介绍的方法可能找不到最佳的可能性量表。我们将每条题项都视为对现有量表的最佳可能补充，但是这里介绍的方法并没有完全探索所有可能的题项组合。在执行定序累计量表分析的软件中，使用者可以定义其他的题项组合，或者作为扩展的起点，或者作为最终的量表。但是在现实中很少这么做。

第2节 不在量表中的题项:拒绝和排除

这套寻找程序最严格的要求是所有的题项对必须是正相关的,即具有正向的 **H**(ij)值。这就是说,当一些题项加入了量表,与它们负相关的任何一个题项就不可能再成为同一量表的一部分了。在聚类法的每一步我们都要检查量表中的每个题项和不在量表中的其他题项之间的 **H**(ij)系数是否是负值。如果量表外的题项和量表中的题项之间的同质性系数是负的,那么这些量表外的题项就要被拒绝。

如果严格执行这一方法,可能会导致拒绝一些这样的题项:它们具有一个稍微负的 **H**(ij)值,如−0.01,但是它们与其他题项之间的 **H**(ij)系数和它们自身的 **H**(i)系数非常高。如何取舍就要看研究者自己的判断:是丢弃该题项,还是把它包括到量表分析中作为起点或者测试(不是寻找)程序的一部分。研究者有时倾向于更有原则性(拒绝题项),有时又会有更加实际的考量(包括该题项)。

在有些情况下,即使题项所有的同质性系数都是正向的,也不应该纳入量表,比如当 **H**(ij)系数和之后的 **H**(i)系数过低(通常低于0.30)时。这类的题项叫做被排除在量表之外。

在有些情形下，一个特别的潜变量可能需要用两个或更多相关概念来理解，而这些概念并不构成一个单一的累计性量表。比如，具有 28 个题项的一般健康问卷（General Health Questionnaire with 28 items，GHQ-28）构成了一个很弱的累计性量表，但是它由四个更强的子量表构成。这四个分量表测量了健康的不同维度，比如心理健康子量表和躯体健康子量表。寻找程序可以在一大组题项中进行，来探索哪些题项属于同一个累计性量表。被第一个量表排除或拒绝的题项可以继续用来看它们是否能构成第二个量表。这种方法可以反复进行，直到没有更多的量表被发现。通常这种方式最终会得到一个包括很多题项的大量表，以及一些剩余的不同的小量表。这些小量表在反映不能够通过大量表测量的潜在特征的方面是很有用的。我们推荐利用这些剩余的量表再进行新的聚类分析，有可能它们就测量到了不同的潜变量。如果这些小量表被用作起始数据集，已经在第一个量表中的题项能够在其他某一个量表中表现得更好。这种方法叫做扩展性寻找。[9]

第5章

一个累计性量表的例子：美国宗教信念

2002 年世界价值观调查的美国数据集包括了五个有关基督宗教信念的问题:“你是否相信……1.……地狱;2.……死后重生;3.……天堂;4.……人有灵魂;5.……上帝。”[10] 我们想要知道这五个问题是否可以作为测量宗教信念程度的单一累计性量表的好指标。人们对这五个问题给出正向回答(“是”)的比例是不同的。正如 2002 年世界价值观调查中的其他国家,相信天堂的美国人比相信地狱的要多。但是这些问题是相关的吗?对于不同的人它们的含义是不同的吗?如果这些问题是相关的,那么相信地狱也意味着相信天堂,反之亦然。但是实际情况真的如此吗?表 5.1 给出了没有加

表 5.1 五个宗教信念问题的回答情况(%) ($N = 1\,200$)

你是否相信	是	否	不知道
地　　狱	858(71.5%)	286(23.8%)	56(4.7%)
死后重生	909(75.8%)	207(17.2%)	84(7.0%)
天　　堂	1 018(84.8%)	138(11.5%)	44(3.7%)
人有灵魂	1 129(94.1%)	48(4.0%)	23(1.9%)
上　　帝	1 133(94.4%)	48(4.0%)	19(1.6%)

权过的问题和答案。问题是按照给出肯定回答的人数比例来排列的。“你相信地狱吗?”是难度最大的题项，给出的正向回答也最少，而“你相信上帝吗?”是最容易的题项。如果这五个题项构成了累计性量表，那么就是按照这种顺序。

第1节 | 寻找相关的基础信息

首先让我们对数据进行二分,把“不知道”看做一种负向回答。[11]五个题项会产生5(5−1)/2=10个交互表。对于每一个交互表,我们可以计算多少个回答者违反了累计模型(Err(obs)),多少个回答者在两个题项处于统计独立性条件下会违反累计模型(Err(exp)),以及这对题项的同质性系数。表5.2给出了计算方法。“死后重生”简写为“重生”,“相信地狱”简写为“地狱”。

表5.2 重生(横向),地狱(纵向)

	0	1	总 数
0	157	*134*	291
1	185	724	909
总 数	342	858	1 200
Err(obs) = 134	Err(exp) = 208.1		**H** = 0.36

对表5.2的解释:909个回答者对“重生”题项给出了正向回答(1=是),858个回答者对“地狱”题项给出了正向回答。因此“地狱”题项比“重生”题项的难度大,普及度低。模型违反被定义为回答者对难度高的题项(“地狱”题项)给出正向回答,对较容易的题项(“重生”题项)给出负向回答。总

共有134个回答者属于此范围,因此Err(obs)=134。相应单元格的值用斜体表示。如果这两道题项("地狱"和"重生")是统计上不相关的,那么交互表的(0, 1)单元格的值应该为(291*858)/1 200=208.1,因此Err(exp)=208.1。这两个题项构成的题项对的同质性系数为**H**=1−Err(obs)/Err(exp),或者**H**(地狱,重生)=1−134/208.1=1−0.64=0.36。

对于这一个交互表,我们用较容易的题项表示行变量,用较难的题项表示列变量。这种方式产生的交互表的右上角单元格(0, 1)表示模型违反的数目。如果行变量是较难的题项,列变量是较容易的题项,那么模型违反的数量则是左下角(1, 0)单元格的值。

第2节 | 总结必要的基本信息

判断模型拟合所需要的信息——Err(obs)，Err(exp)，以及 $\mathbf{H}(ij)$——可以由三个半角矩阵来总结，或者下三角矩阵，如表5.3和表5.4所示：

表5.3 下三角表示每个题项对的模型违反数目实际值，上三角表示每个题项对的模型违反数目期望值

	地 狱	重 生	天 堂	灵 魂	上 帝
地 狱		208.1	130.1	50.8	47.9
重 生	134		137.9	53.8	50.8
天 堂	5	66		60.2	56.8
灵 魂	7	11	17		63.0
上 帝	6	19	8	38	

表5.4 每个题项对的同质性系数

	地 狱	重 生	天 堂	灵 魂	上 帝
地 狱					
重 生	0.36				
天 堂	0.96	0.52			
灵 魂	0.86	0.80	0.72		
上 帝	0.87	0.63	0.86	0.40	

第 3 节　计算单个题项和整个量表的同质性

基于每对题项的基本信息，我们计算每个题项的同质性系数：$\mathbf{H}(i)=1-\mathrm{Err(obs)}/\mathrm{Err(exp)}$，其中 Err(obs)和 Err(exp)分别是包括该题项的所有题项对的实际观测到的和预期的误差总数。

Err(obs)	Err(exp)		$\mathbf{H}(i)$
天　堂	134＋5＋7＋6＝152	208.1＋130.1＋50.8＋47.9＝436.9	0.65
重　生	134＋66＋11＋19＝230	208.1＋137.9＋53.8＋50.8＝450.6	0.49
天　堂	5＋66＋17＋8＝96	130.1＋137.9＋60.2＋56.8＝385.0	0.75
灵　魂	7＋11＋17＋38＝73	50.8＋53.8＋60.2＋63.0＝227.8	0.68
上　帝	6＋19＋8＋38＝71	47.9＋50.8＋56.8＋63.0＝218.5	0.68

我们也可以用同样的方式计算整个量表的同质性系数。

整个量表：Err(obs)		Err(exp)
152＋230＋96＋73＋71＝622		436.9＋450.6＋385.0＋227.8＋218.5＝1 718.8
$\mathbf{H}=1-622/1\,718.8=0.64$		

请注意，此处我们用了两次误差数目的观测值和期望值。把这两个数字除以 2 不会改变我们的结果。

第 4 节 | 统计显著性检验

如果一个量表的 **H** 值是 0.64,那么该量表就是一个很好的累计性量表。但是我们仍然不知道这个量表有多大的可能性是从一个统计独立的数据中产生的(参阅附录第 1 部分)。我们用 $Z(ij)$ 代表一个题项对的值,$Z(i)$ 代表单个题项的值,Z 代表整个量表的值。因为检验题项是否构成量表使用单侧检验[我们比较 $\mathbf{H}>0$ 和 $\mathbf{H}\leqslant 0$(单侧),而不是 $\mathbf{H}=0$ 和 $\mathbf{H}\neq 0$(双侧)],我们可以定义一个单侧的显著性水平 alpha,比如 5%(对应的 z 值为 1.64),2.5%($z=1.96$),1%($z=2.33$),或者 0.1%($z=3.10$)。

因为所有的 $Z(ij)$ 值都高于 3(表 5.5),很显然所有的 10 对题项对的同质性都是具有显著性的。我们也可以对每一个题项分别计算 $Z(i)$ 值:$Z_{地狱}=28.07$,$Z_{重生}=22.36$,$Z_{天堂}=$

表 5.5 $Z(ij)$:对统计独立性零模型进行检验

	地狱	重生	天堂	灵魂	上帝
地　狱					
重　生	11.05				
天　堂	22.30	13.49			
灵　魂	11.86	12.21	14.74		
上　帝	11.67	9.31	17.11	13.34	

33.02，$Z_{灵魂}=25.01$，$Z_{上帝}=24.52$。最后我们也可以计算整个量表的 Z 值：$Z=41.39$。所有的 $Z(i)$值和整个量表的 Z 值都是高度显著的。

这种方法背后的逻辑很清楚，但是实际应用却很有限。即使 alpha 从 0.05 调整到 0.003 1，Z 的临界值（只）增加到 2.73，而 $Z(i)$值增加到超过 20，Z 则超过 40，造成了 $\mathbf{H}(i)$和 $\mathbf{H}$ 的超级显著性。

第 5 节 | 使用成对的信息寻找最佳的量表

既然所有的 $\mathbf{H}(ij)$ 都高于 0.30,那么很自然地所有的五个题项整体构成了一个可以接受的累计性量表。但是这种情况并不永远成立。经常出现能够构成量表的题项对的同质性系数并不全部大于 0.30,或者并没有先验的理由去判断所有的题项都是潜变量的有用指标。在这样的情况下,研究者可以使用分层级自下而上的聚类分析方法,其中同质性系数被用作相似性系数。

第 4 章介绍并解释了这种方法。现在我们用这五个题

表 5.6 寻找程序的简要总结

题项	均值	第一步		第二步		第三步		第四步	
		$\mathbf{H}(i)$	$Z(i)$	$\mathbf{H}(i)$	$Z(i)$	$\mathbf{H}(i)$	$Z(i)$	$\mathbf{H}(i)$	$Z(i)$
地狱	0.72	0.96	22.30	0.94	25.07	0.92	27.68	0.65	28.07
天堂	0.85	0.96	22.30	0.93	27.63	0.88	31.27	0.75	33.02
上帝	0.94			0.87	19.78	0.69	23.36	0.68	24.52
灵魂	0.94					0.64	22.08	0.68	25.10
重生	0.76							0.49	22.36
量表		0.96	22.30	0.92	29.78	0.80	37.06	0.64	41.39
拒绝的题项		无		无		无		无	

项作一简要演示。首先选择最佳的题项对:题项对("地狱","天堂"),**H**(ij)=0.96。第三个加入的题项是"上帝",它与题项对("地狱","天堂")具有最高的 **H**(ij)。第四个题项是"灵魂",第五个也是最后一个题项是"重生"。表 5.6 简要地展示了这一寻找方法。

第6节 | 使用转置的二分数据矩阵

正如所有研究对象都给出相同回答的题项不能够用来计算同质性系数,对所有题项都给出相同回答的研究对象不能够用来对转置数据矩阵进行同质性分析。在美国人宗教信念的例子中,超过70%的被访者相信地狱的存在,这道题项是难度最高的。对所有五个题项都回答"是"的被访者有716人(60%),对所有五个题项都回答"否"的被访者有28人(2%)。对整个量表的同质性系数的计算只能基于剩下的456人。表5.7给出了结果。$\mathbf{H}^T$的值是0.47,是比较满意的,只有2.6%的人具有负的同质性$\mathbf{H}^T(s)$。一般的原则是,如果具有负值的同质性系数的比例低于10%,我们就认为这一数据集是可以构成量表的。

没有给出差异性回答的研究对象没有包括到同质性的计算中,但是如果这些题项构成一个量表,那些研究对象仍然具有一个量表值。对所有题项全部给出负向回答的研究对象的量表值最低(0),对所有题项(k个)全部给出正向回答的研究对象的量表值最高(k)。

表 5.7　转置的数据矩阵的同质性系数以及研究对象的同质性系数

整个量表的 $H^T = 0.47$；	值为负的 $H^T(s) = 12(2+2+2+5+1)$ $H^T(s)$百分比为 2.6%；		剩余研究对象数量为 456
$H^T(s)$的分布			O 表示最多 15 个人
频数	$H^T(s)$的范围		
	<	<=	
	0.9	1.0	|
46	0.8	0.9	|OOOO
74	0.7	0.8	|OOOOO
	0.6	0.7	|
114	0.5	0.6	|OOOOOOOO
59	0.4	0.5	|OOOO
129	0.3	0.4	|OOOOOOOOO
1	0.2	0.3	|O
	0.1	0.2	|
21	0.0	0.1	|OO
2	−0.1	0.0	|O
2	−0.2	−0.1	|O
	−0.3	−0.2	|
2	−0.4	−0.3	|O
5	−0.5	−0.4	|O
	−0.6	−0.5	|
	−0.7	−0.6	|
1	−0.8	−0.7	|O
	−0.9	−0.8	|
			|
28	极　值	0	|OO
716	极　值	1	|OOOOOOOOOOOOOOOOO |OOOOOOOOOOOOOOOOO OOOOOOOOOOOOOO

第7节 | 使用新建立的量表的参数

确定以上五个题项构成了一个量表之后,下一个问题就是:“我们可以利用这一信息来做什么?”每一个统计模型,包括每一个测量模型,都有两个衡量标准。第一,该模型是否拟合我们分析的数据?第二,相应的参数是什么?我们如何来理解它们?我们关于同质性的讨论是有关模型拟合的讨论。我们在下一章会继续探讨更多的关于模型拟合的内容,但是同质性分析的结果显示我们的数据拟合定序 IRT 模型。

如何理解相应的参数?因为模型是定序模型,因此参数是有排序的:题项有排序,研究对象也有排序。题项的排序很容易理解:题项按照普及度(或容易程度)递减来排列,通常就是样本中题项的普及度。研究对象则是按照在五个题项上的总分来排序。相关信息在表5.8中列出。因为这个模型的参数不是在定距水平上来测量的,该模型有时也被称为非参数模型。这个术语有时让人困惑,因为参数都是被估计出来的。因此更好的描述这一模型的方法应该是非数值或者定序模型。

五个题项总分的中位数是5,平均值是4.2,标准差为1.2。根据 t 检验,女性在宗教信念量表上的得分(4.3)比男性(4.1)稍高,具有统计显著性。毫不意外地,当我们比较研究

对象在“你对教会有多大信心?”这一问题上的回答时,回答“非常有信心”或者“很有信心”的研究对象在量表上的平均总分要显著高于回答“不是很有信心”的研究对象,前者为4.6和4.4,后者只有3.5。回答“不是很有信心”的研究对象的平均分又要显著高于回答“一点都没有信心”的研究对象的平均分(2.8),当然后者的数量很少。注意到我们在此处是在定距而不是定序的意义上使用量表得分的。第9章我们会讨论这种解释的合理性。

表5.8 宗教信念的量表得分

得分	频　数	百分比
0	28	2
1	39	3
2	71	6
3	98	8
4	248	21
5	716	60
	1 200	100%

第 8 节 | 结论

在本章中,我们以世界价值观调查中五个询问美国人宗教信念的题项为例,来说明题项的可量测性。所有的题项均可以构成量表:每个题项既有正向回答也有负向回答,所有题项的同质性系数都在统计上显著并且具有足够大的系数值以作合理的解释。不到 40%的研究对象是对某些题项正向回答而对其他题项作出负向回答。在这些研究对象中,只有很小比例的人具有负的同质性系数。研究对象的这些测试值可以在以后的分析中继续使用。

到目前为止,我们的重点在于如何理解对构成哥特曼量表的理想回答模式的偏离。心理测量学家把对同质性系数的强调称为用误差理论来测试确定性模型。现代的 IRT 模型的重点已经从确定性模型转移到了概率性模型,后者用概率的方法来理解每个回答而不是对回答作正确和错误的分类。下一章我们就转到介绍此类概率性模型。

第6章

概率性支配模型：单调同质性

第1节 不理想的回答还是概率性的回答?

我们先回到第2章介绍的确定性累计性量表模型。在一个确定性量表中,在难度较高的题项上给出正向回答的研究对象必然对所有较容易的题项也给出了正向回答。在这样的量表中不存在误差或者模型违反。但如果我们允许一定数量误差的存在,如同在第3章到第5章中所做的,即使我们把误差控制在某个界限以下,我们也相当于重新定义了该模型,它这时已经不再是确定性模型了。那么这样的模型是什么性质的呢?

由于影响回答的因素太多,很多还是未知的,因此期待研究对象给出的回答都符合理想的哥特曼量表是不实际的。在本章中我们寻求一个更加合理的做法。我们不再假设具有特定量表值(θ_s)的研究对象总是对低于该量表值($\delta_i < \theta_s$)的题项作出正向回答,而只是假设具有较高量表值的研究对象有更大——或至少不低于——的概率对具有较低量表值的某一题项作出肯定的回答。因此我们转换术语:我们不再讨论“误差”或者“模型违反”,而只讨论“回答倾向”和“回答概率”。每种回答模式都是可能的,有一定的发生概率,不再意味着它可能包括了一个或多个误差。

确定性的累计性量表模型也能够用概率来表示:当$\delta_i > \theta_s$时,给出正向回答的概率是0,但当$\delta_i \leqslant \theta_s$时,概率则上升

为1(图6.1)。这个概率函数叫做题项回答函数(item response function，IRF)。[12]在确定性模型中，IRF被称为阶梯函数，因为正向回答的概率直接从0上升为1，而不取任何中间值。在一个确定性量表中，不同题项的IRF只是在题项的量表值$\theta_i\delta_i$的位置上有区别。

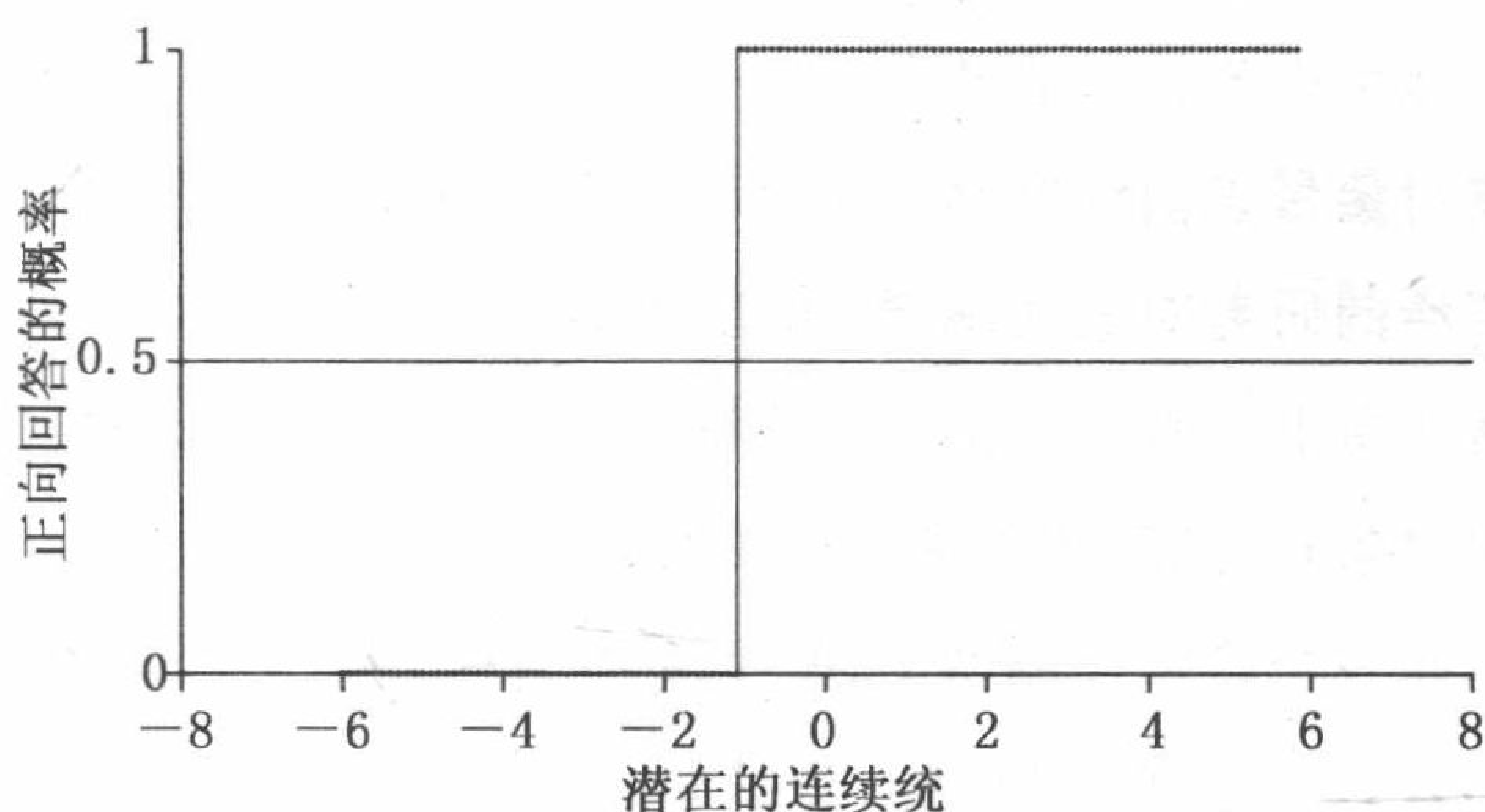


图6.1　哥特曼量表:题项是函数

当回答模式违反了确定性模型时，这些概率就不是0或1，而可能是中间的任何一个值。如果采用一种更宽泛的标准来定义累计性量表，我们则可以要求对某一题项正向回答的概率随着个人能力的增加而增加(至少不会减少)，如图6.2

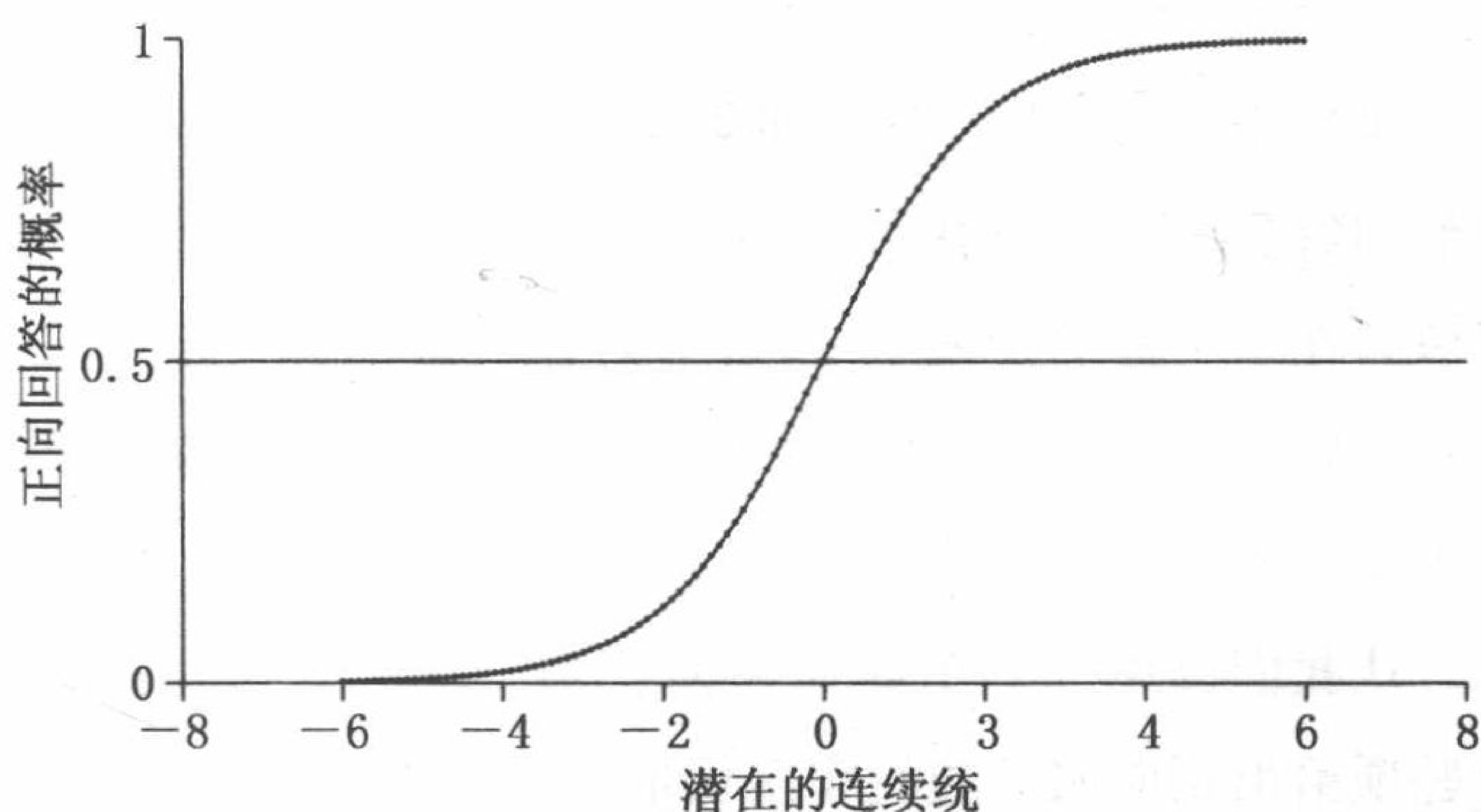


图6.2　题项回答概率函数

所示。在这样的概率模型中,IRF 看起来更接近于 S 形的曲线(与 logistic 函数类似)。

递增的 IRF 函数

概率性累计量表模型的一个最低要求是 IRF 不会随着研究对象量表值的升高而下降。也就是说,在潜变量上具有特定值的研究对象对题项 **i** 做出正向回答的概率至少应该和在潜变量上具有较小值的每个研究对象相同(如能力较低的研究对象)。满足该要求的题项被称为单调同质性题项。

在一个确定性模型中,某个题项的量表值可以被定义为当具有 θ_s 量表值的研究对象从负向回答变为正向回答时在潜在连续统上的相应数值,即 $\delta_i=\theta_s$。在概率性模型中,某题项的量表值与研究对象 **s** 的量表值相同,对 **s** 而言做出正向回答的概率是 0.50。

局部随机独立性

第二个要求是一个隐性的要求,它要求某个体对一个题项的回答只取决于该研究对象在所要测量的潜在特征上的取值。在一个概率性模型中,每个研究对象都有一定的概率对某一题项做正向回答,这个概率基于其在量表上的位置,而不是受到其他系统性的因素影响。

让我们设想一下个体如何回答两个题项。如果他对每个题项给出正向回答的概率仅仅取决于他的量表值,那么他对两个题项均正向回答的概率就是两个独立概率的乘积。

比如,如果研究对象A正向回答题项i的概率是0.3,正向回答题项j的概率是0.4,那么他同时正向回答题项i和j的概率则为0.3×0.4=0.12。研究对象B的量表值稍高于研究对象A,正向回答题项i和题项j的概率分别为0.5和0.6,那么他同时对两个题项都做出正向回答的概率为0.5×0.6=0.30。

要求某个体正向回答某题项的概率仅取决于该个体在潜变量上的取值这一要求叫作局部随机独立性。概率性累计量表模型假设这一要求是被满足的。

总结一下,概率性支配模型有三个模型假设,类似于第2章介绍的决定性模型的假设:

1. 要测量的特征是一个单一的特征,并且可以由一个单一的连续统来表示。

2. 正向回答某题项的概率不随着研究对象在潜变量上取值的增加而减少。

3. 正向回答某题项的概率仅仅由研究对象在潜变量上的取值决定,而不受其他任何系统性因素的影响(即局部随机独立性假设)。

第2节 | 两个概率性模型：单调同质性和双重单调性

当一个累计性量表满足了前述的三个假设，我们就可以在潜在的连续统上对研究对象进行排序，即，在一个潜变量上对他们进行定序测量。这种模型叫做单调同质性模型(monotone homogeneity, MH)，或者单一单调性模型。它是莫坎创立的两种模型中的第一种。

第二种模型的限制性要强一些。莫坎加入了一个假设，即对所有题项做出正向回答的概率的顺序对于所有的研究对象都是一样的，无论它们在潜变量上的取值如何。这样每个研究对象 **s** 对在潜在维度上取特定值的题项有一个正向回答的概率，这一概率比研究对象正向回答量表值更高的题项的概率要大(或者至少一样大)，因为量表值更高的题项的难度更高，普及度更低。这一更具限制性的模型叫做双重单调性模型(double monotonicity, DM)，因为在这种模型中，不仅每个题项的 IRF 概率随着研究对象量表值的升高而单调递增(即图 6.3 中从左到右的趋势)，对每个研究对象而言，他们正向回答题项的概率随着题项量表值的增加而降低(即图6.3 中从上往下的趋势)。

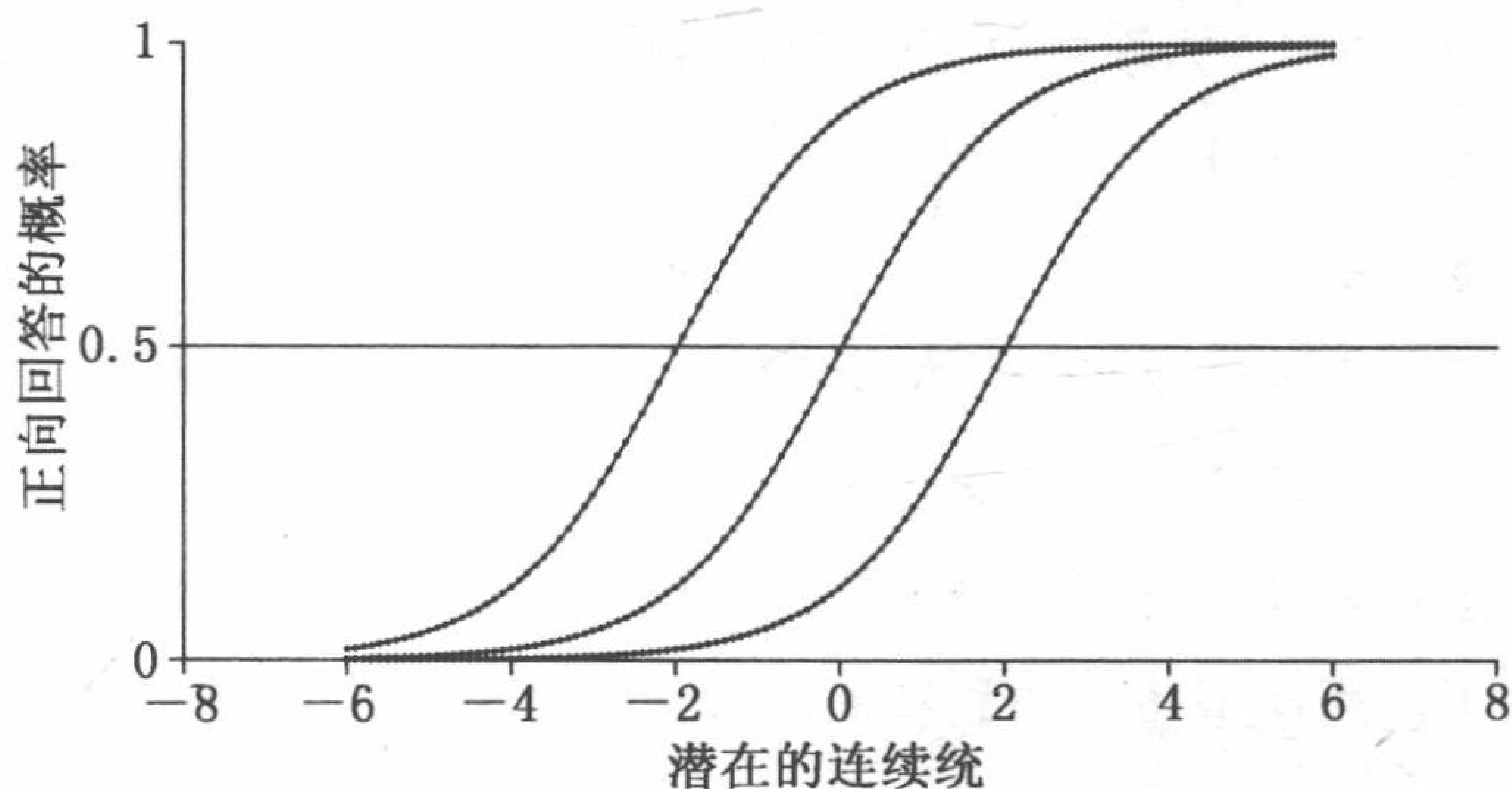


图 6.3　三个双重单调题项

这一假设意味着所有的研究对象在题项难度的排序上都是一致的,无论他们各自的量表值是多少。如果 IRF 之间没有任何相交,那么题项难度的排序在概率上而言对于所有的研究对象都是一样的。图 6.4 给出了一种违反该假设的情况:对于 $\theta < 0.5$ 的研究对象而言,题项 C 是最难的(在此处题项 C 的 IRF 与题项 B 的 IRF 相交);对于 $0.5 < \theta < 1.5$ 的

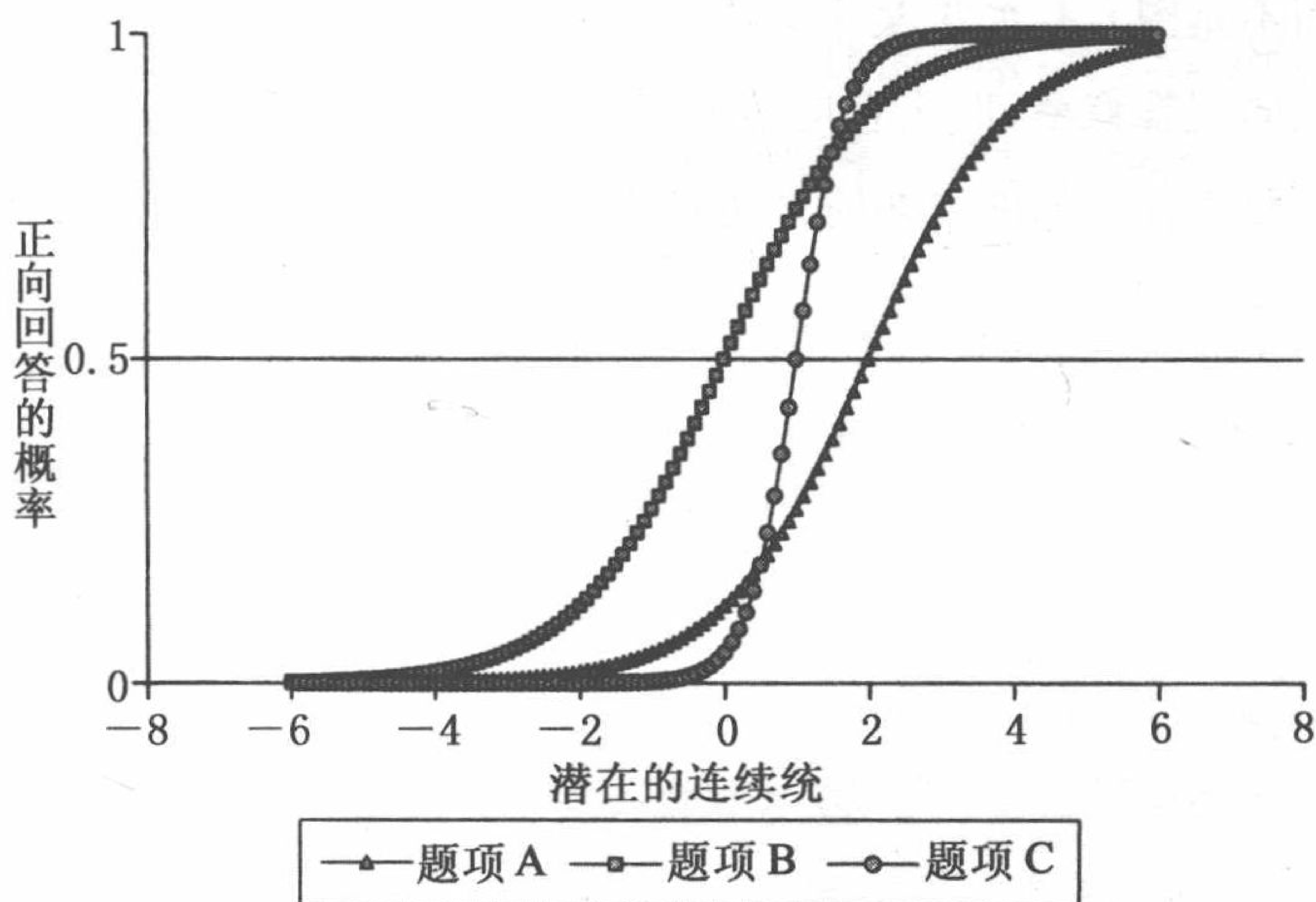


图 6.4　三个非双重单调的题项

研究对象而言，题项 C 的难度居中（在此处题项 C 的 IRF 与题项 A 的 IRF 相交）；对于 $\theta > 1.5$ 的研究对象而言题项 C 的难度是最低的。

可以这样来理解 DM 模型的重要性。在第 3 章我们已经用违反传递关系的概念来定义模型违反，并且据此来检验某个量表及其题项在非理想累计性量表模型中的同质性。为了使用模型违反的数量来检验一个备选量表，我们必须知道题项难度的排序。只有当所有研究对象在题项难度排序上都是一致的，我们才能判断哪些研究对象在特定的题项对中正向回答了难度高的题项，但却负向回答了难度低的题项。如果难度低和难度高的题项对于不同的个体的意义不同，我们无法做出每个题项对的交互表，也无法判定误差单元格。因此我们假设题项难度的排序对于所有的研究对象都相同，稍后我们会检验该假设。

如果我们比较不同题项的 IRF，它们应该如图 6.3 所示而不是图 6.4：如果某个题项对某研究对象而言具有更高的正向回答概率，那么它应该对所有研究对象都如此，不论研究对象各自的能力如何。换句话说，题项的 IRF 不应该相互交叉。

第 3 节 | 检验单调同质性模型

这里我会介绍一种对单个题项的 MH 模型的检验，比如题项 **k**，该题项属于 n 个在某种程度上违反了 MH 模型的可量测二分(0, 1)题项的一个。我们可以利用从量表中其他 $(n-1)$ 个题项的信息重新建构题项 **k** 的 IRF，然后检验对题项 **k** 的正向回答的概率是否随着研究对象量表值的增加而增加(至少不会下降)。

研究对象的量表值建立在其余 $n-1$ 个题项的总分之上，范围是 0 到 $n-1$。得分相同的研究对象属于同一个分数组。这样的分组叫做余分分组，因为量表分数值是由其余的题项决定的。如果题项 **k** 满足 MH 的要求，那么余分值高的研究对象对题项 **k** 正向回答的概率应该也比较高。下面用一个例子来对该原则做一说明。

我们现在有七个题项，每个的回答分类为(0, 1)，构成一个累计性量表。只有一个题项 **k** 的 **H**(i)系数较低，为 0.20，因此不满足 MH 的模型要求(见图 6.5)。

每一个研究对象都要回答题项 **k** 以及其他题项。表 6.1 列出了每个余分分组(根据其他题项的总分来划分)对题项 **k** 的回答情况。注意，这七个题项的普及度比较低，因此大多数研究对象都回答 0。只有 36 个研究对象得到了最高分 7

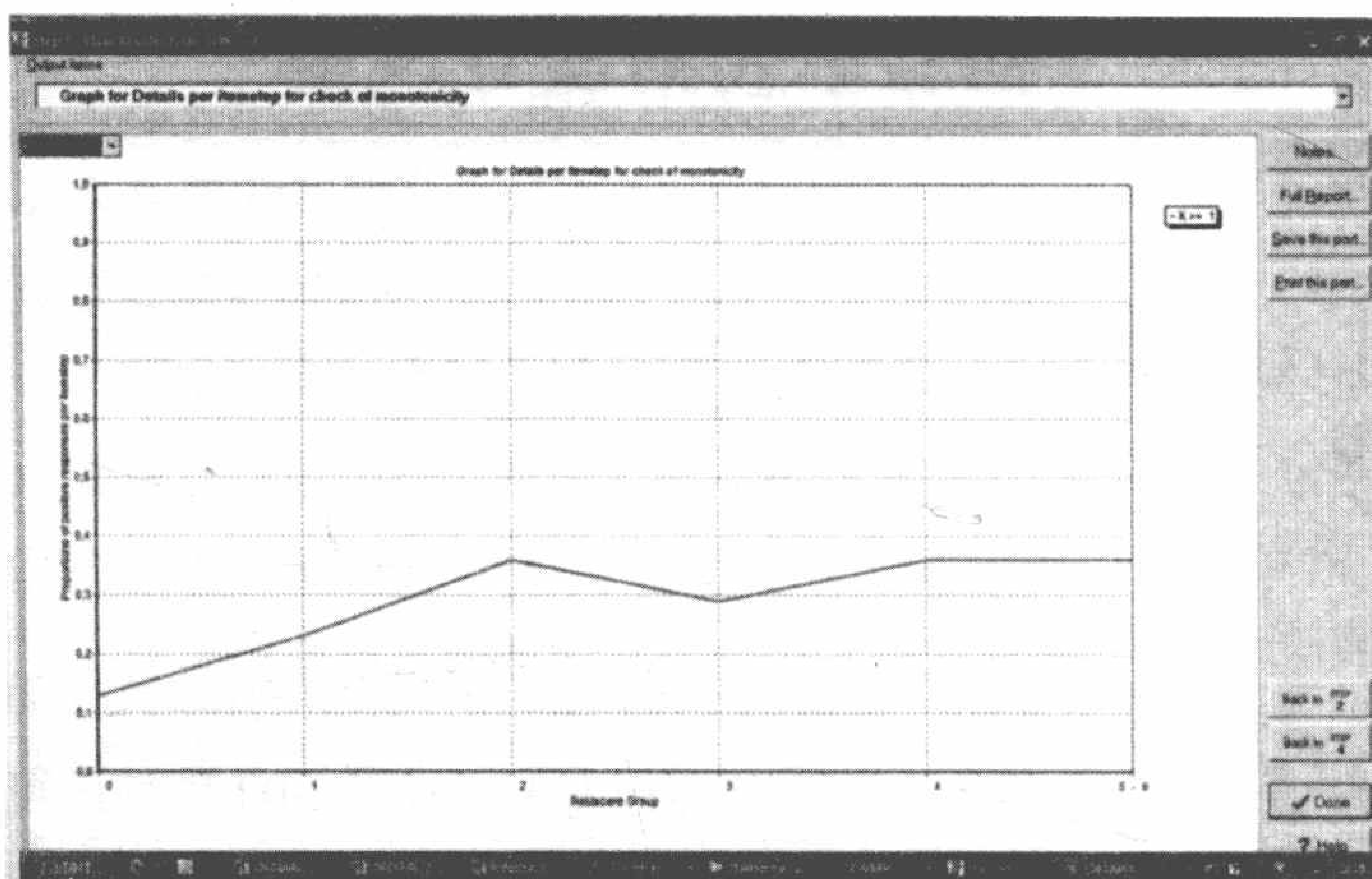


图 6.5 余分分组示意图(分数值处于 0 至 5—6 之间)及题项 k 的正向回答比例,一个违反(分数值处于 2 至 3 之间)

分。我们可以把这群人看做一个小群体,或者可以把他们与前面的组合并以构成一个规模大些的新的组 6A。

研究对象对题项 **k** 正向回答的比例在第一组是 0.13(505/3 885),在第二组中上升至 0.23(244/1 054),在第三组中上升到 0.36,然后在第四组中略微下降到 0.29。之后在第五组中重新增至 0.36,在第六组中则下降到 0.34,最后在第七组中再次增加到 0.44。

如果把最后两组合并为组 6A,那么正向回答的比例仍然为 0.36。因为每组的人数最好是在 50 以上,这样每个研究对象在组中的比例不会超过 2%,我们忽略了回答比例从第五组到第六组的下降。忽略这一违反的另一原因是下降的规模很小,只有 0.02(0.36−0.34)。原则上来讲,只有规模在 0.03 或以上的违反才有研究的意义,但这也取决于研究者的决定。

表 6.1　关于题项 k 是否违反了单调同质性模型的详细信息

余分分组						
低	高	N	每一题项值的频次		平均分	每一题项步骤的正向回答比例
			0	1		≥1
1	0	3 885	3 380	505	0.13	0.13
2	1	1 054	810	244	0.23	0.23
3	2	840	538	302	0.36	0.36
4	3	838	591	247	0.29	0.29
5	4	253	162	91	0.36	0.36
6	5	163	107	56	0.34	0.34
7	6	36	20	16	0.44	0.44
[6A	5 和 6	199	127	72	0.36	0.36]

我们重点关注第三组和第四组间的模型违反,并且假设在合并第六组和第七组后只有六个不同的分组(见图 6.5)。值为 0.064 8(0.359 5－0.294 7)的模型违反的程度有多严重?它可以被余分分组规模从 840 变为 838 的随机波动所解释吗?

表 6.2　检验第三和第四组是否来自同一总体的交互表

	题项 **k** = 0	题项 **k** = 1	总　数
第三组	538	302	840
第四组	591	247	838
总　数	1 129	549	1 678

如表 6.2 所示,在 2×2 表中从 0.359 5(302/840)到 0.294 7(247/838)的下降在两个总体构成相同的条件下可能由抽样变异导致。超出的概率值是第四组及题项 **k**＝1 的对应单

元格人数在超几何分布边缘值为 840，838，1 129 和 549 的条件下取值 247 或更少的概率(参考 Molenaar & Sijtsma，2000：71—72)。

超几何分布正态近似的正态离差 $z=2\times[\sqrt{(f_{11}+1)\times(f_{00}+1)}-\sqrt{(f_{01}\times f_{10})}]/\sqrt{(N+1)}$，或者 $z=2\times[\sqrt{(248\times 539)}-\sqrt{(302\times 591)}]/\sqrt{1\,679}=2\times(\sqrt{133\,672}-\sqrt{178\,482})/40.98=2\times(365.61-422.47)/40.98=2.78$。

在显著度为 5%的单侧检验下，这个 z 值显著大于 1.64 的临界值。因此，这个违反在统计上是显著的，它并不是由抽样误差所导致的。

这种显著的模型违反足以让我们放弃题项 **k** 吗？穆伦纳尔和赛斯玛(Molenaar & Sijtsma，2000)设计了一个标准，叫作 crit，考虑到很多模型违反的其他方面：题项的数量，或者叫作余分分组中与其他题项比较的数目[合适的比较(ac)]，这些比较中违反的数量(# vi)，以不同形式表现出的违反的强度[比如最大规模的违反(maxvi)，违反的总量(sum)，统计上显著的违反数量(# sig)]，以及低 **H**(i)系数(低于 0.30)。crit 的正式定义为：

$$\text{Crit}=50[0.30-\mathbf{H}(i)]+\sqrt{\#\text{vi}}+100\#\text{vi}/\#\text{ac}+100\text{maxvi}+10\sqrt{\text{sum}}+1\,000\text{sum}/\#\text{ac}+5\text{zmax}+10\sqrt{\#\text{sig}}+100\#\text{zsig}/\text{ac}$$

模型违反可以由大于 0.30 的同质性系数 **H**(i)来弥补。在这种情况下，0.20 的低 **H**(i)系数加到该标准中。MH 的要求适用于所有的分组配对。如果有 7 个不同的余分分组，那么就有 21 个分组配对或者合适的比较。6 个分组则只有 15 个分组配对。与正向回答概率为 0 或 1 的极值组的比较

是不必要的，那么比较的数量在这种情况下就会减少。如果只有一个分组配对出现了模型违反的情况，就像上面这个例子，那么这个违反就是最大违反和违反总数量。因为该违反在统计上显著，所以（最大的）显著 z 值是 2.78，并且只有这一个违反是显著的（# zsig）。

这样 crit 值就等于：$50\times0.10+\sqrt{1}+100\times1/15+100\times0.06+10\sqrt{0.06}+1\,000\times0.06/15+5\times2.78+10\sqrt{1}+100\times1/15=5+1+6.7+6+2.5+4+13.9+10+6.5=55.6$。

穆伦纳尔认为低于 40 的 crit 值可以被随机波动来解释，高于 80 的 crit 值就构成了严重的题项。这里的 crit 值为 56，处于两者中间，需要研究者进一步的注意。

第4节 检验五个宗教信念题项的单调同质性

实际检验每个题项的IRF是否单调递增是通过比较研究对象在其他题项上的总得分进行的,即余分。表6.3给出了研究对象在五个题项上的总分,及每次删掉一个题项的五组余分。

表6.3 按总分划分的每组的研究对象人数,及除去某题项后每个余分分组的研究对象人数

正向回答数量	所有5道题项	不包括				
		地狱	重生	天堂	灵魂	上帝
0	28	28	28	28	46	49
1	39	41	53	48	80	76
2	71	75	117	107	102	106
3	98	226	157	301	254	250
4	248	830	845	716	718	719
5	716					
总数	1 200	1 200	1 200	1 200	1 200	1 200

对于五组余分分组的每一组,表6.4列出了被删除的题项的正向回答的频次,用$N(1)$来表示,以及用频次除以组规模的比例(p)。对于"地狱"这道题项而言,第一组:0/28 =

0.00;第二组:2/41=0.05;第三组:6/75=0.08;第四组:134/226=0.59;第五组:716/830=0.86。这些比例的次序的确是单调递增的:0.00—0.05—0.08—0.59—0.86。IRF单调递增的性质在所有五个题项中都有体现(见表6.4)。

表6.4 不同余分分组中研究对象对删掉题项正向回答的数量和比例

组号	余分	地狱		重生		天堂		灵魂		上帝	
		N(1)	p	N(1)	p	N(1)	p	N(1)	p	N(1)	p
1	0	0	0.00	0	0.00	0	0.00	18	0.39	21	0.43
2	1	2	0.05	14	0.26	9	0.19	59	0.74	58	0.76
3	2	6	0.08	60	0.51	45	0.42	90	0.88	93	0.88
4	3	134	0.59	119	0.76	248	0.82	246	0.97	245	0.98
5	4	716	0.86	716	0.85	716	1.00	716	1.00	716	1.00

因为我们有5个不同的分数组,因此可以做5(5—1)/2=10次不同的组与组的比较(第一组对第二组,第一组对第三组……第四组对第五组)。在每次的比较中,次序低的组的比例不应高于次序高的组。然而,如果比例为极值0或极值1,那么相应的比较就完全可以预测:没有一个组的比例是低于0或者大于1的。这意味着比例为极值0或极值1的组会从比较中排除。因此,对于"灵魂"和"上帝"题项来说,所有的10次比较仍是有用的,但是对于"地狱"和"重生"题项来说,只有6次比较是有用的,对于"天堂"题项只有3次有用的比较(第二组对第三组,第二组对第四组,第三组对第四组)。

在这个例子中每个组的规模都是最小值25,并且所有的5个组都是不同的。但是由于有些组会比较小,通常我们需要规模足够大的分组。如果我们把分组的规模提高到至少120个研究对象,MH测试会是什么样子?表6.5重新汇总了表6.3中的

信息。对于“地狱”题项来说,前三个组需要合并来得到120个研究对象。“上帝”题项需要合并前两个组,或者第二组和第三组,来达到120个研究对象的水平。现在表6.6列出了如果其中一个题项被删除后,相应的每一个组 p 值的单调递增情况。

表6.5 按总分划分的每组的研究对象人数,及除去某题项后每个余分分组的研究对象人数(最小组规模:120人)

	不包括									
	地狱		重生		天堂		灵魂		上帝	
组	a	b	a	b	a	b	a	b	a	b
1	0—2	144	0—2	198	0—2	183	0—1	126	0—1	125
2	3	226	3	157	3	301	2—3	356	2—3	356
3	4	830	4	845	4	716	4	718	4	719
		1 200		1 200		1 200		1 200		1 200

注:a. 余分值。
b. 有余分值的研究对象人数。

表6.6 不同余分分组中研究对象对删掉题项正向回答的数量[$N(1)$]和比例(p)

	删掉的题项									
	地狱		重生		天堂		灵魂		上帝	
组	$N(1)$	p	$N(1)$	p	$N(1)$	p	$N(1)$	p	$N(1)$	p
1	8	0.06	74	0.37	54	0.30	77	0.61	79	0.63
2	134	0.59	119	0.76	248	0.82	336	0.94	338	0.95
3	716	0.86	716	0.85	716	1.00	716	1.00	716	1.00

除了“天堂”题项只能比较第一组和第二组,其他的题项都可以做三次比较(第一组对第二组,第一组对第三组,第二组对第三组)。第三组在“灵魂”题项和“上帝”题项上的 p 值接近于而不完全等于1.00(分别为716/718和716/719)。所有的题项都满足MH模型的要求。

第7章

概率性支配模型：双重单调性

第1节 双重单调性的重要意义

我们很难夸大MH模型的重要性。因为我们使用的测量量表没有一个标准的测量单位,并且我们找的测量值通常取决于所测量的特定研究对象群体,所以在实际应用中很难比较来自不同文化、不同样本或者不同时期的测量。符合DM模型要求的测量量表则意味着在任意一个总体、任意一个样本、任意一个时期,或者任意一个实验设计下,基于量表值对研究对象进行排序得到的结果都是一样的。

我们将讨论评价DM模型的五种不同方式。第一种已经在第3章和第5章讨论过了,即使用转置的数据矩阵。在评价量表和单个研究对象的同质性时,我们对所有研究对象使用同一个题项次序。如果研究对象的同质性很高,我们就可以推论这种假设的模型违反不严重。除此之外,本章会介绍四种其他的模型检验方法:使用外部群体,余分方法,合并余分方法,以及P(+, +)和P(−, −)矩阵。

第 2 节 | 使用外部群体测试双重单调性

如果所有题项难度的排序对于所有的研究对象都一样，那么我们可以做的就是检验这一顺序是否也适用于与研究相关的研究对象的子群体。这些子群体的划分可以基于人口学变量,比如性别、年龄、教育程度、教会成员资格,或者态度变量,例如在前面提到的美国人的宗教信念的例子中,划分子群体的标准是祷告的重要性或者对教会的信心程度。具体的标准取决于不同的研究情境。

世界价值观调查还询问了关于宗教信仰的其他问题,比如“宗教信仰在你的生活中有多重要?”和“你对教会的信心程度如何?”在未加权的美国人数据集中,454 个研究对象对教会非常有信心,450 个研究对象对教会很有信心,其余的 296 个研究对象则没有那么大的信心,或者完全没有,或者没有回答。

我们可能有疑问:这三组研究对象在考虑前述五道测量宗教信念程度潜在特性的题项上是否存在差异? 这些题项对于不同的三组人的意义是一致的吗? 如果答案是否定的，那么题项普及度的排序很可能在三组人中是不同的。如果真是这样,实际上相当于研究对象对不同的题项做出了正向回答,而我们却将他们的量表值放在一起比较。表 7.1 给出

了对三组研究对象的同质性分析结果。

这三组研究对象在题项同质性上是存在差异的。但是我们首要关心的是这是否还是同一个量表,也就是说,其中的量表是否具有相同的普及度排序。这是我们要研究的DM模型的要求。然而我们发现实际情况不是这样的:对于"非常有信心"的组,"重生"题项比"地狱"题项的普及度要低(0.81比上0.86),对于"没有很多信心"的组来讲,"灵魂"题项比"上帝"题项的普及度要高(0.85 比上 0.83)(这些比例在表7.1中斜体表示)。

表 7.1 宗教信念量表的 p 值和 H 值,以及数据集和分组关于教会信心程度的题项回答情况

			你对教会的信心程度如何?					
	总样本		非常有信心		很有信心		不是很有或根本没有信心	
	$p(i)$	$\mathbf{H}(i)$	$p(i)$	$\mathbf{H}(i)$	$p(i)$	$\mathbf{H}(i)$	$p(i)$	$\mathbf{H}(i)$
地狱	0.71	0.65	*0.86*	0.32	0.72	0.63	0.47	0.75
重生	0.76	0.49	*0.81*	0.25	0.81	0.48	0.59	0.51
天堂	0.85	0.75	0.96	0.65	0.89	0.72	0.61	0.71
灵魂	0.94	0.68	0.97	0.56	0.97	0.65	*0.85*	0.65
上帝	0.94	0.68	0.99	0.50	0.97	0.63	*0.83*	0.63
N, **H**	1 200	0.64	454	0.38	450	0.61	296	0.65

我们可以在每个分组中根据题项的普及度来对题项进行排序。因此在"非常有信心"的组中我们调换了"重生"和"地狱"题项的顺序,在"不是很有信心或根本没有信心"的分组中调换了"上帝"和"灵魂"题项的顺序。这些顺序的调换带来了更高的同质性系数值(表 7.2)。最后一组("不是很有或者根本没有信心")的差别可以忽略,但是第一组的变化还

是非常明显的。

表 7.2　在分组内部调整题项后的宗教信念量表的 p 值和 H 值

你对教会的信心程度如何？					
非常有信心			不是很有或根本没有信心		
	$p(i)$	$\mathbf{H}(i)$		$p(i)$	$\mathbf{H}(i)$
重生	*0.81*	0.32	地狱	0.47	0.75
地狱	*0.86*	0.41	重生	0.59	0.51
天堂	0.96	0.65	天堂	0.61	0.71
灵魂	0.97	0.56	上帝	*0.83*	0.65
上帝	0.99	0.50	灵魂	*0.85*	0.68
N, **H**	454	0.45		296	0.66

那么我们如何检验题项排序的差别是否可以由在单一总体中的随机变异来解释呢？让我们用“非常有信心”的组在“重生”和“地狱”题项上的回答举例说明(表 7.3)。

表 7.3　“对教会很有信心”分组中“重生”题项(纵向)和“地狱”题项(横向)的交互表

	地　　狱		总　　数
重生	0	1	
0	21	63	84(19%)
1	41	329	370(81%)
总数	62(14%)	392(86%)	454

在这个分组中，“重生”题项比“地狱”题项的普及度要低，因为(0, 1)单元格比(1, 0)单元格的值要大(63 比上 41)。关于这些差异可以被单一总体的随机变异所解释的 McNemar 检验由穆伦纳尔提出(Molenaar, 1970：100，公式 5.5)。他用下面这一公式来表示：

$$z = \sqrt{(2k+2+b)} - \sqrt{(2n-2k+b)}$$

$$\text{其中 } b = [(2k+1-n)^2 - 10n]/12n$$

其中

在这个公式中 k 是两个频次中最小的一个,41; n 是两个频次的总和: $41+63=104$。

$$\text{所以 } b = [(2\times 41+1-104)^2 - 10\times 104]/12\times 104$$
$$= [(-21)^2 - 1\,040]/1\,248 = (441-1\,040)/1\,248$$
$$= -0.48$$

$$z = \sqrt{(2\times 41+2-0.48)} - \sqrt{(2\times 104-2\times 41-0.48)}$$
$$= \sqrt{83.52} - \sqrt{125.52} = 9.139 - 11.204 = 2.065$$

这个值是显著的,我们需要认真考虑这一结果。

这一结果是否应该让我们放弃"重生"题项或者"地狱"题项也取决于模型违反的其他方面,比如 crit 值。"地狱"题项的 crit 值是 41,"重生"题项的 crit 值是 49。在这个例子中我们应该同时接受这两个题项,即使它们存在显著的模型违反题项。

在对分组进行量表分析时还应注意两点。首先,题项的普及度在分组与分组之间可能有很大差别。我们可以想象,宗教信念这些题项在对教会信心很高的分组中可能普及度更高,在其他组中则不然。实际上这是真实存在的。在试图确定所有题项是否测量同样一个潜在特性的量表分析中,题项在普及度上有差别是没有问题的。不同分组在量表值上存在的系统差别只是强调了量表的有效性。第二,量表和题项在不同子样本中可能具有不同的同质性系数。这也不构成问题。可能的情形是,454 个对教会具有很大信心的研究

对象可能都具有很强的宗教信念,同时具有相同的高量表值。如果是这种情况,那么这组研究对象的同质性过高,在局部随机独立性的假设下,他们回答中存在的任何差别都可以归因为随机波动。反过来说,这意味着实际观察到的模型违反数量与期望的模型违反数量非常接近,因此同质性系数很低。

同质性系数是一个很灵活的相关系数,并且相关系数的值取决于题项的方差。一组研究对象的同质性越高,在散点图上的点越近似圆圈,而不是长长的雪茄状,同时对该组研究对象拟合回归线也更困难。如果研究对象在某个维度上差别很小,那么就很难找到他们之间的差别了。这实际上是局部随机独立性假设的扩展:如果所有研究对象在潜在维度上具有相同的数值,那么同质性量表是不存在的。因此如果我们想比较子样本,我们很少去比较它们同质性系数的差别。[13]

第3节 | 使用余分分组测试双重单调性

用不同的外部群体来比较题项的顺序是否就足够了?答案是否定的,从来没有一种测试是足够的。这一回答来自波普尔有关科学的证实和证伪的思想。根据波普尔的思想,个人永远无法知道某些事物是永远真实的,因为不可能每时每刻都去测试它(Popper, 1959, 1963/2003)。然而只要没有发现相反的事实证据,接受某事物的暂时真实性是可能的。可能带来反面证据的严格检验并不能推翻某说法的真实性(暂时的)是一个很好的科学思路,比如接受量表 DM 模型的要求。

除了利用外部群体还有其他的方式来检验 IRF 是否相交。我们可以利用一种与检验 MH 的方法类似的方式。针对量表值逐渐增加的研究对象组,我们分别观察每一个题项的 IRF,量表值是基于在其他 $(n-1)$ 个题项上的得分。然后针对量表值逐渐增加的研究对象组,我们观察每一对的 IRF 的难度顺序,这里的量表值是基于其他的 $(n-2)$ 个题项得分。利用 $n-2$ 个题项,我们最多可以区分 $n-1$ 个组,量表值在 0 和 $n-2$ 之间(二分回答,0 和 1)。每组的研究对象都要回答被观察的题项对。我们把题项对用($\mathbf{i}$, $\mathbf{j}$)表示。因此在每一个余分分组($\mathbf{i}$, $\mathbf{j}$)中,有四个可能的回答:00, 01, 10,

11。如果题项 j 比题项 i 在整个样本中更容易或普及度更高,那么题项 j 在每一个余分分组中也要比题项 i 的普及度更高。因此在每一个分组中,$f(01)+f(11)>f(10)+f(11)$,或者更简单地,$f(01)>f(10)$。如果事实并非如此,那么我们就用二项式分布的正态逼近来检验我们是否应该接受总体中 $f(01)=f(10)$ 的假设。与之前利用外部群体检验 DM 的模型违反相同,在这里我们也采用 McNemar 检验。

我们在美国人宗教信念的例子中进行这一检验。首先,我们进行基于余分分组的比较,余分分组是根据五个题项中的三个题项得到的。余分分组的规模很明显取决于哪些题项对被删除。表 7.4 给出了组规模。

表 7.4 删掉两个题项形成的余分分组的研究对象人数

	被删除的题项									
余分	地狱 重生	地狱 天堂	地狱 灵魂	地狱 上帝	重生 天堂	重生 灵魂	重生 上帝	天堂 灵魂	天堂 上帝	灵魂 上帝
0	28	28	46	51	28	58	51	47	57	113
1	58	53	85	77	70	123	142	125	110	111
2	120	239	232	237	256	171	158	309	313	254
3	994	880	837	835	846	848	849	719	720	722

结果显示,10 组中只有 1 组稍微违反了 DM 模型,该余分分组建立在"重生"题项和"天堂"题项被删除的基础上(图 7.1)。对于四个余分分组的每一个我们可以制作关于被删除题项的交互表(表 7.5)。

表 7.5　对基于其余三个题项划分的四个不同余分分组做题项“重生”和题项“天堂”的交互表

组	得分	N	“重生”和“天堂”的交互表				v	z	p 重生	p 天堂
			00	*01*	*10*	*11*				
1	0	28	28	0	0	0			0.00	0.00
2	1	70	39	9	14	8	0.07	0.83	0.31	0.24
3	2	256	48	37	52	119	0.06	1.49	0.67	0.61
4	3	846	1	129	0	716			0.85	1.00
总数		1 200	116	175	66	843	0.13		0.76	0.85

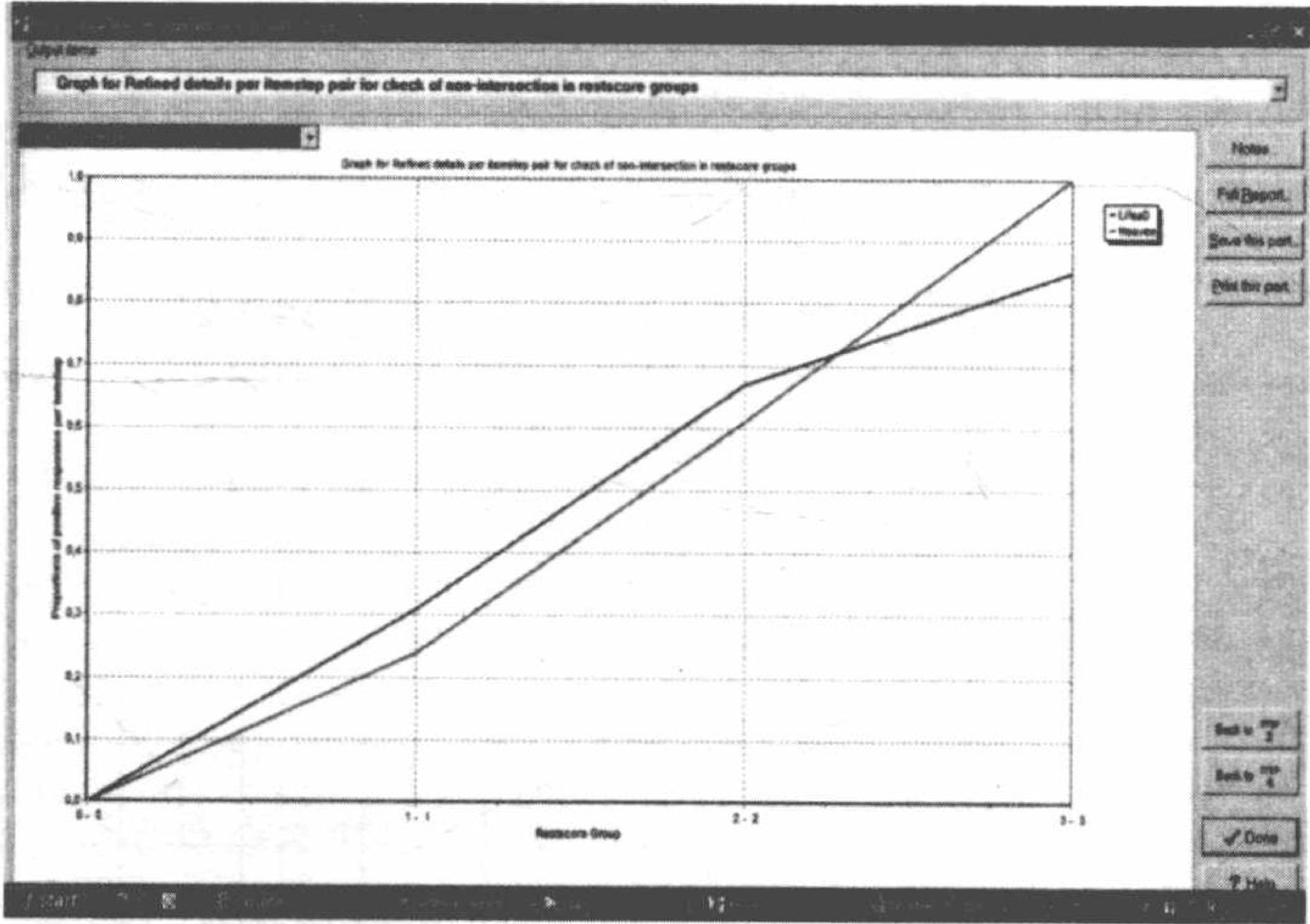


图 7.1　题项对“重生”和“天堂”违反 DM 模型的图示

表 7.5 中的 N 和表 7.4“重生天堂”列下的 N 是相同的。在总体中,“天堂”题项比“重生”题项要更容易,普及度更高。因此“天堂”题项在四个分组中的每一个也应该比“重生”题项容易。在第一个余分分组中,研究对象在其他的三个题项上都给了负向回答(因为总分为 0),没有人对这两个题项给

出正向回答。

最后一个分组中的研究对象对其他的三个题项都是正向回答的，716 个研究对象对这两个题项也是正向回答的。129 个研究对象正向回答了第二个（“天堂”）题项，负向回答了第一个（“重生”）题项（列 01），但是没有人给出相反的回答模式，即（1，0）的回答模式。因此总体来说，846 个研究对象中，716 个正向回答了“重生”题项（比例为 0.85），845（129 + 716）个研究对象正向回答了“天堂”题项（比例为 1.00）。该组中“天堂”题项更高的正向回答比例与总体中的趋势是一致的，因此该分组符合 DM 模型的要求。

然而，第二组和第三组分别在余分 1 和 2 上违反了 DM 模型的要求。对于第二组来说，“重生”题项（$p=0.31$）比“天堂”题项（$p=0.24$）的普及度更高。这一在反方向上的差别（v）是 $0.31-0.24=0.07$。第三组存在相同的题项，“重生”题项（$p=0.67$）比“天堂”题项（$p=0.61$）的普及度更高，差别为 $0.67-0.61=0.06$。

我们应该担心这些模型违反吗？或者它们可以由随机离差所解释吗？因为在这种规模以及边缘频次分布的群体中的随机离差很容易导致模型违反。答案存在于之前使用外部群体来解释的检验中，相关的 z 值分别是 0.83 和 1.49。两者都低于显著度为 5%的单侧检验的临界值 1.64，因此这些违反在统计上都不显著。此外，crit（重生）= 37，crit（天堂）= 24，均低于正常值 40，因此我们不需要过分担心。

第4节 使用合并余分分组测试双重单调性

余分分组方法可能存在的一个问题是组规模比较小，由此造成检验的显著度不够强。为了提高显著度，我们可以对不同的余分分组进行重新划分。如果有分数在0到$m-1$之间的m个分组，那么就存在$m-1$种方式把它们合并到更低或更高的组中：(0分组)比上(1到$m-1$分组)；(0和1分组)比上(2到$m-1$分组)……(0到$m-2$分组)比上($m-1$分组)。对于$m-1$个二分类中的每一种，我们现在可以同时在低分和高分余分分组检验DM假设。

我们可以对表7.4中的四个不同余分分组(分数0—3)应用这一原则，用三种不同的方式对它们进行合并或二分(表7.6、表7.7和表7.8)。表7.6中的第一个二分方式没有

表7.6 第一种二分：0分组比上1—3分组

组	得分	N	“重生”和“天堂”的交互表				v	z	p 重生	p 天堂
			00	*01*	*10*	*11*				
1低	0—0	28	28	0	0	0			0.00	0.00
2高	1—3	1 172	88	175	66	843			0.78	0.87
总数		1 200	116	175	66	843			0.76	0.85

发现模型违反,但是后面的两个二分就出现了违反现象(表7.7和表7.8)。第一个违反不显著(McNemar检验中 $z=0.83$),但是第二个就具有统计显著性 ($z=1.80$)。注意,两个违反都是在低分组中发生的,而不是在高分组中。

表7.7　第二种二分:0—1分组比上2—3分组

组	得分	N	"重生"和"天堂"的交互表				v	z	p 重生	p 天堂
			00	*01*	*10*	*11*				
1低	0—1	98	67	9	14	8	0.05	0.83	*0.22*	*0.17*
2高	2—3	1 102	49	166	52	835			0.80	0.91
总数		1 200	116	175	66	843	0.05		0.76	0.85

表7.8　第三种二分:0—2分组比上3分组

组	得分	N	"重生"和"天堂"的交互表				v	z	p 重生	p 天堂
			00	*01*	*10*	*11*				
1低	0—2	354	115	46	66	127	0.06	1.80	*0.55*	*0.49*
2高	3—3	846	1	129	0	716			0.85	1.00
总数		1 200	116	175	66	843	0.06		0.76	0.85

我们可以用题项的区分度——题项回答函数的倾斜度——来解释这一现象。如果违反是发生在低分组的,难度较大的题项对于量表分较低的研究对象而言过于简单,对于量表分较高的研究对象则不是如此。总体而言,与其他题项相比这样的题项的斜率不那么陡峭,因此区分度也较少。如果违反是发生在高分组的,难度较大的题项对于量表分较高的研究对象而言过于简单,对于量表分较低的研究对象则不是如此。这样的题项的斜率比较大,区分度也更大。在这个例子中,"重生"题项似乎比"天堂"题项的区分度更小。

第5节 使用P(+,+)和P(-,-)矩阵测试双重单调性

另外一种检验IRF是否相交的方法是遵循局部随机独立性假设的逻辑。假设有三个题项 **i**, **j** 和 **k**,按照普及度的排序为 $p(i) > p(j) > p(k)$。在IRF不相交的条件下,这一顺序对每个研究对象 s 都适用,同时量表值 θ_s 满足 $p^s(i) > p^s(j) > p^s(k)$ 的条件。局部随机独立性假设意味着对于每个研究对象 s 来说,对一系列题项的回答概率等于对每个题项的回答的概率的乘积。所以对于研究对象 s 对题项 **i** 和 **j** 都给出正向回答的概率为 $p^s(ij) = p^s(i) \times p^s(j)$。这意味着,在局部随机独立性的假设下存在 $p^s(ij) > p^s(ik) > p^s(jk)$。如果IRF不相交,这一顺序适用于所有研究对象以及整个样本,所以我们可以删除上标 s: $p(ij) > p(ik) > p(jk)$。正向回答(+)的信息被储存在一个方阵中,叫做P(+,+)矩阵。当矩阵中的题项按照普及度从低到高的顺序排列时,每一行和每一列的单元格的频次应该按照比例从小到大递增。

对于负向回答(-)的逻辑是一样的:在局部独立性假设成立且IRF不会相交的条件下,当 $1-p(i) < 1-p(j) < 1-p(k)$,或者 $p(i)^- < p(j)^- < p(k)^-$ 时,可以推出 $p(i^- j^-) <$

$p(i^- k^-) < p(j^- k^-)$。这些信息被存在 P(－，－)矩阵中。当矩阵中的题项按照普及度从低到高的顺序排列时，每一行和每一列的单元格的频次应该按照比例从高到低递减。这里我会证明，如果一个特定的模型违反发生在 P(＋，＋)矩阵中，P(－，－)矩阵则不会发生同样的情况，反之亦然。

我们用实质例子来说明这一检验方法。从表 5.1 到表 5.3 我们可以得到 10 个题项对中(1，1)或(＋，＋)单元格的值，以及(0，0)或(－，－)单元格的值。如果在矩阵中的题项按照难度递减的顺序排列，单元格的值对于 P(＋，＋)矩阵应该是单调递增的，对于 P(－，－)矩阵应该是递减的，无论横向还是纵向。我们用频次除以总人数得到比例，而不使用频次。相关的信息已在表 7.9、表 7.10 和表 7.11 中列出。

我们怎样从这些矩阵中看出模型违反的情况呢？因为这些矩阵是对称的，所以无论我们是横向看还是纵向看结果都是一样的。我们就按照纵向来看，从 P(＋，＋)矩阵和“地狱”、“重生”列开始。从这两列往下，我们比较其余三个题项的单元格的值：“天堂”，0.71－0.70；“灵魂”，0.71－0.75；“上帝”，0.71－0.74。因为“地狱”比“重生”的普及度低，所以对这三对数字来说，第一个不应该高于第二个。然而对于“天堂”这行却不是如此，因为 0.71＞0.70。

如果我们只关注“地狱”题项，我们可以把它与其余每一个题项(列)合并，然后按行比较单元格来看一下是否第一个单元格的值高于第二个。对于题项“重生”和“天堂”来说没有问题，但是“灵魂—上帝”这对题项存在不止一个问题：在“重生”那一行，单元格的值是 0.75－0.74。

因为每个题项都与其余 $k-1=4$ 个题项相关联，每对题

项都与 $k-2=3$ 个行题项相关联,所以对每个题项的比较总数是 $(k-1)(k-2)=12$。当第一个比例恰好为 0 或者最后一个比例恰好为 1 时,这个数字可能会变小。

表 7.9 P(+, +)和 P(−, −)矩阵的信息。下三角矩阵:(1, 1)单元格频次;上三角矩阵:(0, 0)单元格频次;$N=1\,200$

	地狱	重生	天堂	灵魂	上帝
地 狱		157	177	64	61
重 生	724		116	60	48
天 堂	853	843		54	59
灵 魂	851	898	1 001		29
上 帝	852	890	1 010	1 091	

表 7.10 P(+, +)矩阵:对称矩阵,根据表 7.9 中下三角矩阵所计算,除以 $N=1\,200$

		地狱	重生	天堂	灵魂	上帝
	p	0.71	0.76	0.85	0.94	0.94
地 狱	0.71		0.60	0.71	0.71	0.71
重 生	0.76	0.60		0.70	0.75	0.74
天 堂	0.85	0.71	0.70		0.83	0.84
灵 魂	0.94	0.71	0.75	0.83		0.91
上 帝	0.94	0.71	0.74	0.84	0.91	

表 7.11 P(−, −)矩阵:对称矩阵,根据表 7.9 中上三角矩阵所计算,除以 $N=1\,200$

		地狱	重生	天堂	灵魂	上帝
	$1-p$	0.28	0.24	0.15	0.06	0.06
地 狱	0.28		0.13	0.15	0.05	0.05
重 生	0.24	0.13		0.10	0.05	0.04
天 堂	0.15	0.15	0.10		0.04	0.05
灵 魂	0.06	0.05	0.05	0.04		0.02
上 帝	0.06	0.05	0.04	0.05	0.02	

现在我们同样再来看 P(－，－)矩阵，其中对于每个题项的成列配对而言，每行单元格的值不应该增加。对于("重生"，"天堂")这对列来说，"地狱"这一行违反了模型设定，从 0.13 增加到了 0.15，以及"上帝"这一行，从 0.04 增加到了 0.05。还有("灵魂"，"上帝")这对列在"天堂"这一行，从 0.04 增加到了 0.05。

如何来评价 P(＋，＋)或 P(－，－)矩阵中的模型违反情况？P(－，－)矩阵中的模型违反告诉我们，虽然"天堂"题项的负向回答比"重生"题项的负向回答的普及度低，但是，对题项对("地狱"，"天堂")的负向回答比题项对("地狱"，"重生")的负向回答的普及度高。所以当我们按照对"地狱"题项的不同回答来比较"天堂"和"重生"的交互表时，肯定有某些地方出错了。对于 P(－，－)矩阵来说，我们尤其关注对"地狱"题项的负向回答，因为这张表的边缘分布给出了题项对("地狱"，"重生")和("地狱"，"天堂")的负向回答的比例。

由于在表 7.12 中，左边(1，0)单元格的值(44)比(0，1)单元格的值小(64)，这也表明了存在模型违反的情况(表 7.12 右边部分可以被忽略)。如果这些值都相同，那么不存在模型违反。我们依然可以用同一测试(McNemar 检验)来检验 44 和 64 两个值是否来自对这两个单元格具有相同回答

表 7.12　"天堂"(横向)、"重生"(纵向)和"地狱"(第三个维度)的交互表

天堂→ 重生↓	地狱 = 0			天堂→ 重生↓	地狱 = 1		
	0	1	总数		0	1	总数
0	113	64	177	0	3	2	5
1	44	121	165	1	131	722	853
总数	157	185	342		134	724	858

概率的同一总体。在这个例子中,z 值为 1.83,是显著的。然而 crit(重生)=31, crit(天堂)=18,因此我们不需要过分担心这一结果。而且模型违反只是 0.02,在默认的检验设置下(最低违反为 0.03),这一违反不会构成太大问题。

第6节 | 从概率性模型的测试中我们可以学到什么?

这些概率性模型的检验通常不是单独进行的,而是针对已经具有足够高同质性的一组题项进行的(参阅 Mokken, Lewis, & Sijtsma, 1996; Roskam, van den Wollenberg, & Jansen, 1986)。它们通常不会拒绝某个量表作为一个具有同质性的测量工具,但是可能会指出是否把某个题项纳入量表还需要再考虑。那么哪一个检验是最重要的?我再次提及证伪的思想:所有的检验都很重要!任何一个显著的证伪都应该引起研究者的注意。我们并没有原则来指导研究者如何挑选外部群体,或者如何合并余分分组。是否删除某些题项或者哪些应该删掉的讨论应该建立在研究需要的基础上,统计检验的结果可以帮助研究者进行选择,但是不能仅仅根据统计检验结果来做决定。

此外,DM 的要求只是平行 IRF 的必要条件而不是充分条件,就像当下最流行的参数 IRT 模型之一——拉希模型所要求的那样。

然而,我们必须说莫坎设定的关于同质性系数的 0.30 的最低界限在实际中还是经得起时间考验的。同质性系数越高,量表中的题项就更可能通过概率性模型的检验。

因此,基于定序 IRT,对于一组题项是否构成一个累计性量表有三个要求:这些题项应该有足够高的同质性——如第 3 章讨论的;IRF 应该单调递增到一个足够的程度(单调同质性检验)——如第 6 章提到的;所有研究对象在题项排序上都是一致的(双重单调性检验)——如我们之前所描述的。

第8章

多分类题项的累计测量

在之前的章节中,我们考察的是如何用一组二分题项构建累计性量表,每个题项把潜在的连续统分为两个区域。对题项做出负向回答的研究对象由连续统的低分端来表示,做出正向回答的研究对象由连续统的高分端来表示。这两个区域由一个基准值分开。这个基准值的位置代表了该题项的位置,由 δ_i 表示,也是题项的参数。因为一个二分类的题项只有一个基准值,所有它只有一个题项参数。

在本章中,我将介绍如何利用相似的原则用含有三个或更多个定序分类的题项构建量表,这一思想是由穆伦纳尔(Molenaar, 1991)提出的。具有三个或更多个定序类别的题项叫做多分类题项。[14]一个例子是,"你对下列说法是同意还是不同意:'如果人们不想工作,就不用去工作。'(1)非常同意,(2)同意,(3)不同意也不反对,(4)不同意,(5)非常不同意。"[15]这个题项可以被认为是测量工作态度的一个指标。因为答案有五类,所以研究对象由潜在连续统的五个定序区域来代表,区域之间存在四个基准值边界。这四个基准值就是这个题项的四个参数,可以从题项 **i** 开始排序,比如 $\delta_{i12} \leqslant \delta_{i23} \leqslant \delta_{i34} \leqslant \delta_{i45}$。每个研究对象 **s** 仍然可以由一个单一的参数 θ_s 来表示,因为他只能落在量表上的一个区域内。

穆伦纳尔建议把一个基准值作为一个题项梯度(item step)。一个二分题项有一个题项梯度，一个 k 个回答类型的多分类题项有 $k-1$ 个题项梯度。

我们首先考察在理想的确定模型中题项梯度的顺序，之后再考察模型违反的概念，类似第 3 章二分题项的做法，我们使用同质性系数来比较观测到的模型违反数量与统计独立条件下的模型违反期望值。接着我们再介绍在多分类题项构成的累计量表中如何使用概率性模型(MH 和 DM)。最后简要地展示一个应用的案例。

第 1 节 | 多分类题项构成的确定性累计模型的回答模式

让我们回到第 2 章第一个例子,这里的潜变量本身(人们的身高)是很明显的。我们用两个二分类的问题来测量身高:A:“你有 1.70 米高吗?”及 B:“你有 1.80 米高吗?”现在我们把这两个问题合并为一个三分类的问题:“你的身高是多少?是否(0)低于 1.70 米,(1)在 1.70 米和 1.80 米之间,或者(2)高于 1.80 米?”对这个问题的回答与对前面两个二分问题的回答是一样的,也就是说它的两个基准值和前面两个二分类题项的两个基准值是相同的(图 8.1a)。我们再用另外一道关于身高的三分类问题做说明:“你的身高是多少?是否(0)低于 1.65 米,(1)在 1.65 米和 1.75 米之间,或者(2)高于 1.75 米?”同理,这个问题也可以看做是两个二分类问题的合并:A:“你有 1.65 米高吗?”以及 B:“你有 1.75 米高吗?”(图 8.1b)

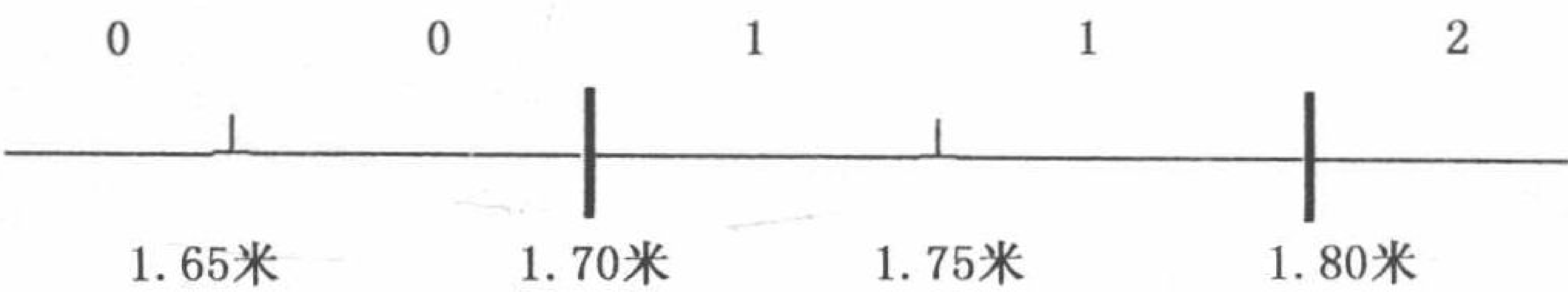


图 8.1a 题项 1:低于 1.70 米(0),高于 1.80 米(2),或者在中间(1)

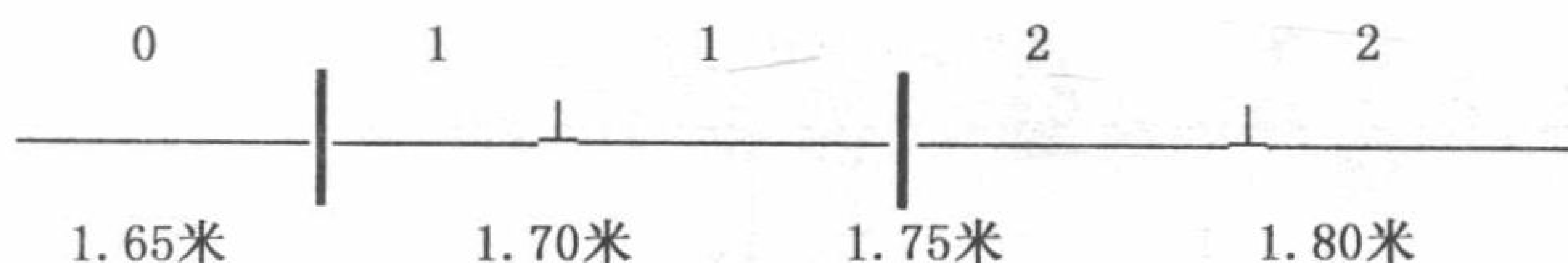


图 8.1b 题项 2:低于 1.65 米(0),高于 1.75 米(2),或者在中间(1)

因为以上两个问题每个都有两个题项梯度,把它们合并起来就得到了四个题项梯度。遵循同样的逻辑,这就相当于四个二分类题项的基准值数目。总起来看,两个三分类题项可以把一组研究对象划分为五个定序的分类。在我们这个例子中,这五个类别是 (0)$s<1.65$ 米;(1)1.65 米$<s<1.70$ 米;(2)1.70 米 $<s<$ 1.75 米;(3)1.75 米 $<s<$ 1.80 米;(4)$s>1.80$ 米。图 8.1c 表示了这一具有四个题项梯度的潜变量。[16]

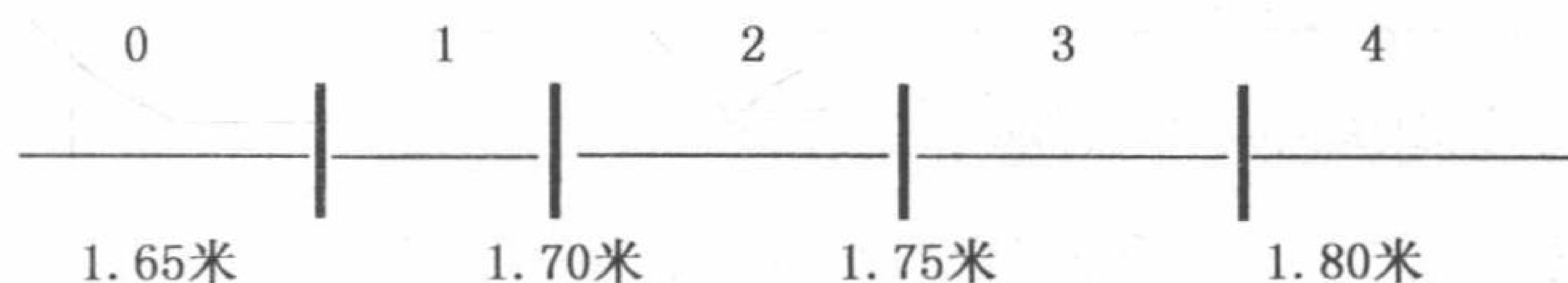


图 8.1c 合并题项 1 与题项 2:从左到右有五个定序类别作为题项梯度之和

在处理多类别题项时,对类别的编码通常从 0 开始进行连续编码。这些编码就表明了潜在连续统上从左往右的题项梯度的数目。通过合并题项(图 8.1c),我们合并了分开的两个题项的题项梯度的数量(图 8.1a 和图 8.1b),来对研究对象的量表值进行排序。这也是我们为什么在图 8.1a 中显示两个 0 和两个 1,在图 8.1b 中显示两个 1 和两个 2。

表 8.1 列出了对这两个问题的回答。逻辑上来讲,九个单元格中只有五个才是有意义的,也就是图 8.1c 所示的潜变量存在的五个区域。

表 8.1 对两个身高题项的可能回答的交互表

题项 1↓ 题项 2→	0：<1.65 米	1：1.65—1.75 米	2：>1.75 米
0：<1.70 米	(0，0)：<1.65 米	(0，1)：1.65—1.70 米	
1：1.70—1.80 米		(1，1)：1.70—1.75 米	(1，2)：1.75—1.80 米
2：>1.80 米			(2，2)：>1.80 米

下面是这个例子中包括了模型违反的四个单元格：四个由于在逻辑上不成立而导致的空单元格：

(0，2)	低于 1.70 米	以及高于 1.75 米
(1，0)	在 1.70 米和 1.80 米之间	以及低于 1.65 米
(2，0)	高于 1.80 米	以及低于 1.65 米
(2，1)	高于 1.80 米	以及在 1.65 米和 1.75 米之间

在构成确定性单维度量表的两个二分题项所构成的 2×2 交互表里，因为它们理想地测量的是同一个潜变量，所以四个单元格里只有一个是空白的。在两个三分类题项构成的理想地测量同一个潜变量的 3×3 交互表中，九个格子中有四个是空白的，在这个例子中是单元格(0，2)、(1，0)、(2，0)和(2，1)。如果把这一理念扩展到社会科学研究中以测量潜在变量，实际上相当于把哥特曼量表的理念一般化了，对多分类题项的所有可能的回答中，只有一部分组合可以进入到哥特曼量表中。

第2节｜在确定性累计量表中使用社会科学题项

现在我们来介绍一个更有趣的例子，读者也许能够自己进行分析。[17]这个例子的数据来自2002年世界价值观调查中的美国样本。这部分数据包括了五道定序类别的问题。这些问题用来测量同一个潜在的行为倾向——政治行为，包括"A.签署请愿书"(请愿)，"B.参与联合抵制"(抵制)，"C.参与合法的游行"(游行)，"D.参与非正式罢工"(罢工)，以及"E.占领建筑物或者工厂"(占领)。

关于这五种行为，每个回答者被询问到：是否曾经做过("曾经做过")，是否可能会做("可能会做")，或者，无论在何种情况下都不会去做("绝对不会做")。我们调换了原来的编码顺序，目的是让三个回答类别按照政治行为从低到高的倾向性排序(0＝绝对不会做，1＝可能会做，2＝曾经做过)。[18]没有做出实质性回答的研究对象将从分析中删除，余下的样本规模$N=1\,110$。表8.2列出了回答情况。

分析多类别题项的一个主要技巧是，把k个回答类别拆分为$k-1$个二分的题项梯度，然后把这些题项梯度作为新的二分题项。对这些题项梯度进行排序和第7章中二分题项的排序方法是一样的，即基于回答的频次。如果研究对象

表 8.2 关于政治行为的五个问题的频次分布表，$N=1\ 110$

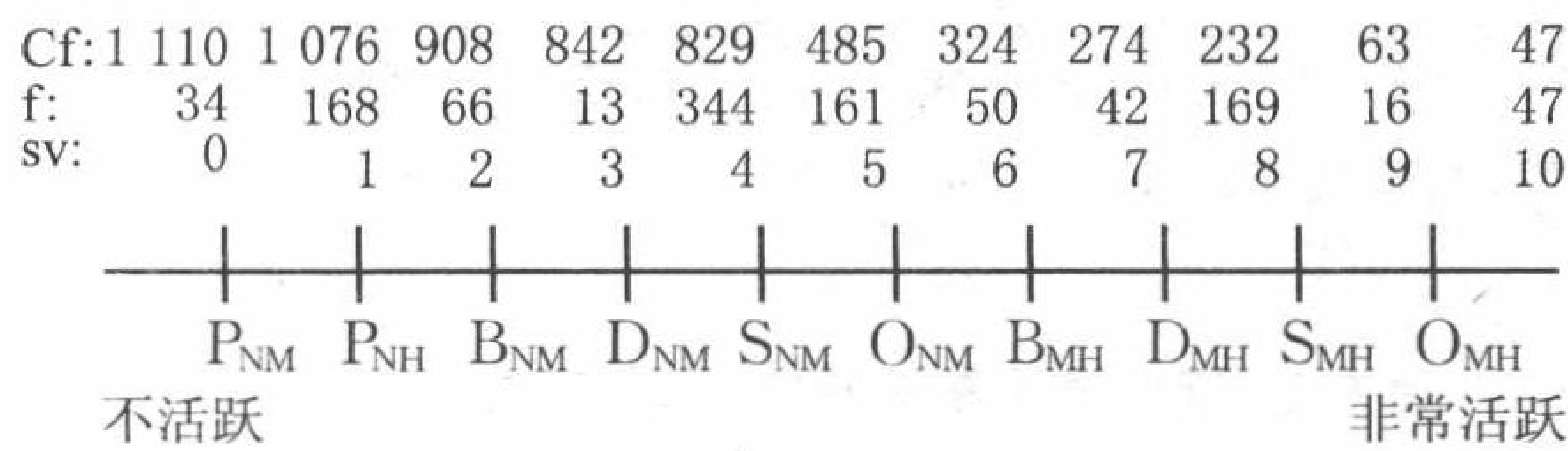
	绝对不会做(0)	可能会做(1)	曾经做过(2)
A. 签署请愿书	34(3.1%)	168(15.1%)	908(81.8%)
B. 参与联合抵制	268(24.1%)	568(51.2%)	274(24.7%)
C. 参与合法的游行	281(25.3%)	597(53.8%)	232(20.9%)
D. 参与非正式罢工	625(56.3%)	422(38.0%)	63(5.7%)
E. 占领建筑物或者工厂	786(70.8%)	277(25.0%)	47(4.2%)

对题项的回答是 m 类别（m 处于最低类别 0 和最高类别 k 之间），那么这可以被看做是对题项梯度 $m-1$, m 做出了正向回答(1)，对于所有选择更高类别的研究对象来说都如此。回答 $m-1$（或更低）的研究对象则被视为对题项梯度 $m-1$, m 做出了负向回答(0)。

表 8.3 给出了对于每个题项的两个题项梯度的回答频次。因为这些题项有三个类别，所以每个题项有两个题项梯度，分别叫作(01)和(12)。(01)代表了对于“绝对不会做”某种政治行为与其他选项（即“可能会做”与“曾经做过”）之间的区分。(12)代表了“绝对不会做”或“可能会做”与其他选项（“曾经做过”）之间的区分。考察这些频次也就是从两个方向考察累计频次，根据负向回答从最低类别到 $m-1$ 类别的频次，以及根据正向回答从 m 类别到 k 类别的频次。根据这些累计频次可以确定题项梯度的顺序。作为二分题项，频次越高，题项梯度的普及度越高，难度越低。如果我们用 P 代表请愿，B 代表联合抵制，D 代表合法游行，S 代表非正式罢工，O 代表占领建筑物，并且用 N 代表“绝对不会做”类别，M 代表“可能做”，H 代表“曾经做过”，那么就可以按照下面的方式对题项梯度进行排序（括号内是正向回答的频次）：

$P_{NM}(1\ 076) - P_{MH}(908) - B_{NM}(842) - D_{NM}(829) - S_{NM}(485) - O_{NM}(324) - B_{MH}(274) - D_{MH}(232) - S_{MH}(63) - O_{MH}(47)$

图 8.2 表示了这一顺序：



注：P_{NM}是题项"请愿"中"从不"和"可能"类别的基准值，Cf 是处于特定题项梯度右侧区域的研究对象的累计频次，f 是处于量表的两个题项梯度之间的特定区域的研究对象人数，sv 代表量表值。

图 8.2 五个题项构成理想的确定性累计量表的条件下，题项 P(请愿)，B(联合抵制)，D(游行)，S(罢工)，以及 O(占领)的基准值

表 8.3 关于政治行为的五个问题的两个题项梯度的频次分布表，N = 1 110

	题项梯度(01)(绝对不会做 N—可能会做 M)		题项梯度(12)(可能会做 M—曾经做过 H)	
	0	1	0	1
A. 签署请愿书(P)	34	1 076	202	908
B. 参与联合抵制(B)	268	842	836	274
C. 参与合法游行(D)	281	829	878	232
D. 参与非正式罢工(S)	625	485	1 047	63
E. 占领建筑物或者工厂(O)	786	324	1 063	47

根据表 8.3 和图 8.2 中的信息，现在可以从这五个题项中重新建构 10 个[5×(5－1)/2]个交互表。表 8.4 给出了题项 B(参与联合抵制)和题项 C(参与合法游行)的交互表的例子。最后一列和最后一行给出了两个题项的累计频次。

(0，0)单元格给出了从量表左侧到题项梯度 B_{NM} 的频次总和 (34＋168＋66＝268)(参见图 8.2)。(1，0)单元格给出了在题项梯度 B_{NM} 和 D_{NM} 之间的研究对象频次(13)。(1，1)单元格给出了 D_{NM} 和 B_{MH} 之间的频次总和 (344 ＋ 161 ＋ 50 ＝ 555)。(2，1)单元格给出了在 B_{MH} 和 D_{MH} 之间的研究对象频次(42)。(2，2)单元格代表了 D_{MH} 右侧剩下的研究对象 (169＋16＋47＝232)。

表 8.4 理想累计量表中两个政治行为题项的交互表

题项 B:参与联合抵制	题项 C:参与合法游行				
	绝对不会做(0)	可能会做(1)	曾经做过(2)	总数	累计
绝对不会做(0)	268 ↓	0	0	268	1 110
可能会做(1)	13 →	555 ↓	0	568	842
曾经做过(2)	0	42 →	232	274	274
总　数	281	597	232	1 110	
累　计	1 110	829	232		

根据这些题项梯度的顺序可以很清楚地看到交互表中的哪些单元格是有数值的,哪些是空白的。这些空白的单元格是误差单元格。这些单元格的实际观测频次和统计独立条件下单元格期望频次之间的比较将决定这些构成基本累计量表的题项对的同质性状况。

任何一个 k 回答类别的题项(通常在 0 到 $k-1$ 之间)具有 $(k-1)$ 个基准值。[19] 因此两个这样的题项总共具有 $2(k-1)$ 个基准值,把潜在连续统划分成 $2(k-1)+1$ 个不同的区域。但是这两个题项存在 $k \times k$ 个可能的回答组合。如果只有 $2(k-1)+1$ 个这样的组合可以被理解为有意义的

量表回答，那么剩下的 $k\times k-2(k-1)-1$ 个则是无意义的，违反了累计模型。随着回答类别 m 递增，可能的模型违反数量也在递增。如果 k 为 2，4 种回答组合里只有 1 次模型违反；如果 k 为 3，9 种回答组合里就有 4 次模型违反；k 为 4，则 16 种回答组合中有 9 次模型违反；k 为 5，则 25 种回答组合中有 16 次模型违反。这些回答组合中哪些是有效的，哪些是模型违反要根据基准值的排序或者题项梯度来断定。

第3节 评价同质性

表8.5给出了用世界价值观调查美国部分的数据做出的两个政治行为问题的交互表。图8.2给出了题项梯度的顺序,因此下面四种回答组合属于模型违反:(0, 1),绝对不会做,可能会做;(0, 2),绝对不会做,曾经做过;(1, 2),可能会做,曾经做过;以及(2, 0),曾经做过,绝对不会做。在交互表里这四个单元格的频次用斜体表示。

表8.5 "参与联合抵制"和"参与合法游行"的交互表

题项B:参与联合抵制	题项C:参与合法游行			
	绝对不会做(0)	可能会做(1)	曾经做过(2)	总数
绝对不会做(0)	149	*101*	*18*	268
可能会做(1)	99	395	*74*	568
曾经做过(2)	*33*	101	140	274
总　　数	281	597	232	1 110

在我们社会科学的例子中,包含了模型违反的回答组合在逻辑上是可能的,而且在经验中也确实会发生。但是它们的确违反了我们用来测量单一潜变量的累计模型。为了更好地评价模型违反的严重程度,或者说缺乏同质性的程度,我们采用与第3章中二分题项相同的做法,把实际观测到的模型违反数量与统计独立条件下模型违反的期望值相比较。

在同质性状况不是很理想的条件下，这些题项无法构成一个累计性量表。

表 8.6 给出了在统计独立性条件下的两个题项的期望频次。这些频次是通过将每个单元格的行边际频次和列边际频次相乘再除以总人数得到的结果。比如，对于(0，0)单元格，对应“绝对不会做”和“绝对不会做”，它的期望频次是 (281×268)/1 110=67.8。与表 8.5 相同，误差单元格的频次用斜体表示。

表 8.6　“参与联合抵制”和“参与合法游行”在统计独立性条件下的期望频次的交互表

题项 B：参与联合抵制	题项 C：参与合法游行			
	绝对不会做(0)	可能会做(1)	曾经做过(2)	总数
绝对不会做(0)	67.8	*144.1*	*56.0*	268
可能会做(1)	143.8	305.5	*118.7*	568
曾经做过(2)	*69.4*	147.4	57.3	274
总　　数	281	597	232	1 110

有四个单元格包括了模型违反，如果这两个问题的答案是相互独立的，单元格中的期望频次分别为：单元格(0，1)：144.1；(0，2)：56.0；(1，2)：118.7；(2，0)：69.4。

分别考察四个误差单元格的误差数量是没用的，我们需要把它们合并起来，总体考察这些题项对的同质性状况。这是因为研究者必须整体考察一个题项——而不是考察它包含的单个题项梯度——是否与量表中的其他题项具有足够高的同质性。最简单的方式就是把四个误差单元格的误差实际观测值总和与统计独立条件下的期望误差数量总和相比较。与第 3 章相同，我们首先讨论一个题项对的同质性，

随后讨论同一个题项对中的单个题项的同质性。

从实际数据交互表(表 8.5)中我们得到实际观测的误差数量 Err(obs)=101+18+74+33=226。以及从统计独立条件下的交互表中(表 8.6)得到统计独立条件下的期望误差数量,Err(exp)=144.1+56.0+118.7+69.4=388.2。与二分题项的例子相同(第 3 章),同质性系数可以由下列公式表示:

$$\mathbf{H}(ij)=1-\frac{\mathrm{Err(obs)}}{\mathrm{Err(exp)}}$$

在这个例子中,$\mathbf{H}(ij)=1-226/388.2=0.42$。在继续确定包括多于两个题项的量表的同质性以及单个题项的同质性之前,让我们进一步考察这个题项对中四个包括模型违反的误差单元格。是否有某些模型违反比其他的违反更严重?比如,单元格(0, 2)似乎比单元格(0, 1)或者(1, 2)的违反状况更加严重,因为它与表 8.4 所展示的理想路径的偏离程度更大。我们应该考虑这种差异吗?穆伦纳尔(Molenaar, 1991)提出通过利用题项梯度的顺序对模型违反进行加权。

加权的同质性系数

我们用两个题项 **i** 和 **j** 做例子,两个都具有三个回答类别(0, 1, 2),假设题项梯度的顺序为 **i**(01), **j**(01), **i**(12), **j**(12)。现在五个有意义的回答组合为(0, 0), (1, 0), (1, 1), (2, 1)和(2, 2)。四个无意义的有误差的回答组合为(0, 1), (0, 2), (1, 2)和(2, 0)(表 8.7)。

表 8.7　交互表，从左(上)到右(下)展示了题项梯度

题项 i↓ 题项 j→	0	1	2
0	(0, 0)	误差	误差
1	(1, 0) ↓→	(1, 1)	误差
2	误差	(2, 1) ↓→	(2, 2)

如果我们对题项梯度进行排序，就可以制作出表 8.8。第一列给出了对两个题项的回答，第一个是对题项 **i** 的回答，第二个是对题项 **j** 的回答。第二列表明了只根据研究对象对题项 **i** 的回答，该研究对象在题项梯度 **i**(0, 1)上是属于左(0)还是右(1)。类似地，第三列、第四列和第五列分别表示了某个研究对象在题项梯度 **j**(01)，**i**(12)和 **j**(12)上是属于左(0)还是右(1)。

表 8.8　题项梯度作为新的二分变量形成的数据矩阵

回答 i, j	i(01)	j(01)	i(12)	j(12)
0, 0	0	0	0	0
1, 0	1	0	0	0
1, 1	1	1	0	0
2, 1	1	1	1	0
2, 2	1	1	1	1

对题项 **i** 回答 0 的研究对象不仅在题项梯度 **i**(01)上是在左边的区域(因此在题项梯度 **i**(01)上被编码为 0)，他/她也必然落在题项梯度 **i**(12)左边的区域，因此在题项梯度 **i**(12)上也被编码为 0。另一方面，对题项 **i** 回答 2 的研究对象在题项梯度 **i**(12)上是在右边的区域，因此被编码为 1，他/她也必然落在题项梯度 **i**(01)右边的区域，因此在题项梯度 **i**(01)上也被编码为 1。这一准则对所有题项都适用。

题项梯度现在可以被看做新的二分变量,并且在难度上有区别:在表 8.8 中的位置越靠右,难度越大。编码 1 和 0 则是对特定题项梯度做出肯定或者否定的回答。根据累计量表模型,对一个难度较大的题项(梯度)做出肯定的回答就意味着对所有难度较低的题项(梯度)也做出肯定的回答。因此,回答组合可以用一个由 1 组成的下三角矩阵和一个由 0 组成的上三角矩阵来表示,就像对于二分题项构成的哥特曼量表一样。

现在我们来考察包括无意义的回答类型组合。

回答类型 **ij**(0, 1)在表 8.9 中被编码为 0—1—0—0。因此较难的题项梯度 **j**(01)得到的是正向回答,较容易的题项梯度 **i**(01)得到的是负向回答。因此根据第 3 章讨论的传递性原则,**i**(01)/**j**(01)这一组合构成了一个模型违反,**ij**(0, 1)就导致了一个误差。回答组合 **ij**(0, 2)的情况更糟,它包含了三个无意义的编码组合,**i**(01)/**j**(01), **i**(01)/**j**(12),以及 **i**(12)/**j**(12),因此导致了三次误差。回答类型组合 **ij**(1, 2)和 **ij**(2, 0)又各自包含一次模型误差,分别为 **i**(12)/**j**(12)和 **j**(01)/**i**(12)。

表 8.9 包括模型违反的新的数据矩阵

回答 i, j	i(01)	j(01)	i(12)	j(12)	误差数目
0, 1	0	1	0	0	1
0, 2	0	1	0	1	3
1, 2	1	1	0	1	1
2, 0	1	0	1	0	1

一个无意义的回答组合所包含的误差数目叫做这个回答组合的权重。为了确定同质性,我们需要用权重分别乘以

观测到的误差数量和期望误差数量。

我们以题项对“联合抵制”和“游行”做例子。从图8.2中可以知道题项梯度的顺序，表8.5则提供了四个误差单元格中每一个的误差频次。表8.10列出了相应的结果。

表8.10 对题项对“联合抵制”—“游行”计算加权的H系数

回答 (B, C)	E(obs)	E(exp)	B_{NM}	D_{NM}	B_{MH}	D_{MH}	W=# errors	W* E(obs)	W* E(exp)
0 0			0	0	0	0	0	0	0
0 1*	101	144.1	0	1	0	0	1	101	144.1
0 2*	18	56.1	0	1	0	1	3	54	168
1 0			1	0	0	0	0		
1 1			1	1	0	0	0		
1 2*	74	118.7	1	1	0	1	1	74	118.7
2 0*	33	69.4	1	0	1	0	1	33	69.4
2 1			1	1	1	0	0		
2 2			1	1	1	1	0		
总数	226	388.2						262	500.2

在这个例子中，加权过的同质性系数（$1-262/500.2=0.48$）甚至比没加过权的系数（$1-226/388.2=0.42$）还要好。注意，误差单元格的权重取决于题项梯度的排序。

量表及单个题项的同质性

一旦理解了题项对的同质性的概念，以及题项梯度在决定模型违反以及其权重中所扮演的角色，进一步考察量表和单个题项的同质性就变得比较简单了，因为其背后的逻辑与第3章对二分题项的讨论是一致的。比如，为了确定整个量表的同质

性，我们需要汇总加权过的实际观测的误差数量和期望误差数量。表 8.11 给出了相关信息。Err(obs)的加权汇总值为 1 544，Err(exp)的加权汇总值为 2 723.9，因此 **H**＝ 1－1 544/2 723.9＝0.43。对于“请愿”题项来说，对应的数字则分别为 Err(obs)＝390 和 Err(exp)＝ 658.3，因此 **H**＝ 1－390/658.3＝ 0.41。

表 8.11 题项(对“联合抵制”,“游行”)。下三角为加权过的观测误差数量；上三角为加权过的期望误差数量

	请愿	联合抵制	游行	罢工	占领
请　愿		237.3	225.6	116.5	78.9
联合抵制	110		500.2	334.0	319.0
游　行	140	262		319.2	295.4
罢　工	65	174	147		297.8
占　领	75	239	178	154	

表 8.12 给出了下三角中成对的 **H**(ij)系数值。对于上三角我们给出了 $Z(ij)$值，对这些值的解释与第 3 章和附录第 1 部分中讨论的相同。只要 $Z(ij)$高于一个临界最低界限，通常是显著度 alpha 水平为 0.05，我们就认为 **H**(ij)系数在统计上是显著的。注意，在这里所有的 **H**(ij)系数值都是正的，整体上来看相当高，题项对(请愿，占领)除外。因为这个 **H**(ij)值比较低(0.05)，所以对应的 $Z(ij)$值也很低(0.44)。

表 8.12 对于每个题项对，下三角中的同质性系数 H(ij)；上三角中的 Z(ij)

	请　愿	联合抵制	游行	罢工	占领
请　愿		11.43	7.91	5.36	0.44
联合抵制	0.54		15.08	11.40	6.20
游　行	0.38	0.48		12.63	9.37
罢　工	0.44	0.48	0.54		12.91
占　领	0.05	0.25	0.40	0.48	

第 4 节 | 多分类题项的寻找程序

与第 4 章的二分类题项相同，多分类题项累计量表的寻找程序也是一种自下而上的方式。由一个最小的可能量表开始（自下），即一个两题项（多分类）构成的量表。在所有的题项对中，我们找到 **H**(ij)系数最高的题项对，即系数值为 0.539 的题项[见表 8.12，该题项对为（游行，罢工）]。题项对（请愿，联合抵制）的 **H**(ij)系数稍微低了一些，为 0.536。这个值是正的，比默认的最低界限 0.30 要高，而且在统计上是显著的[Z(ij)值大于 2]。随着在量表中加入越来越多的题项，量表的同质性和量表中所有单个题项的同质性都会降低，而 Z 值则变大（见表 8.13）。最后一个题项，“占领”，仍然具有一个足够高的 **H**(i)系数，它与题项“请愿”和“联合抵制”构成的题项对的 **H**(ij)系数较低，但是与题项“游行”和“罢工”构成的题项对的 **H**(ij)系数较高，两者刚好抵消。没有方向为负的 **H**(ij)值，所以不存在负的相关系数，因此没有题项被量表拒绝。此外，所有的 **H**(i)值都足够高，因此没有题项被排除。

这一寻找程序的结果在表 8.14 中作为一个最终量表列出。与表 8.13 的区别在于：表 8.14 根据题项的平均值从低到高重新对题项进行了排序。类似于二分题项构成的量表，

表 8.13 分层级聚类(寻找)程序的简要经过

题　　项	平均值	第 1 步		第 2 步		第 3 步		第 4 步	
		H(i)	Z(i)	H(i)	Z(i)	H(i)	Z(i)	H(i)	Z(i)
游　　行	0.96	0.54	12.63	0.50	19.67	0.47	21.10	0.46	23.03
罢　　工	0.49	0.54	12.63	0.51	16.98	0.50	17.60	0.49	21.55
联合抵制	1.01	…	…	0.48	18.84	0.49	22.00	0.44	22.30
请　　愿	1.79	…	…	…	…	0.46	14.48	0.41	13.23
占　　领	0.33	…	…	…	…	…	…	0.35	15.05
量　　表		0.54	12.63	0.49	22.68	0.48	26.86	0.43	30.55
被拒绝的题项		无		无		无		无	

我们可以把这一顺序理解为它们的累计度:在第一个题项("最难")上的值高的话,也意味着在其后所有"更容易"的题项上的值也高(在表中由上而下表示越来越容易)。对于多分类题项而言这只是一个大致的描述,因为我们只是对题项梯度的顺序感兴趣,而不是作为整体的题项。

表 8.14 最终的量表(N = 1 110, H = 0.43, Z = 30.55,临界 Z = 2.73)

题　　项	平均值	题项 H	Z
占领建筑物或工厂	0.33	0.35	15.05
参与非正式罢工	0.49	0.49	21.55
参加合法游行	0.96	0.46	23.03
参加联合抵制	1.01	0.44	22.30
签署请愿书	1.79	0.41	13.23

第5节 对多分类题项应用概率性模型

当我们使用至少包括50个研究对象的余分分组进行MH检验时，结果表明不存在大于0.03的违反。我们展示了一个详细结果，来说明是用哪对余分分组来比较的(表8.15)。首先，得分为0和1的余分分组被合并，得分为7和8的余分分组也被合并，这样使得每个余分组至少有50个研究对象。在一个有三种回答分类的题项中，两个题项梯度被区分出来，(01)和(12)。对于每一个余分分组而言，对题项梯度的正向回答比例表示为题项梯度回答函数。对于第一个余分分组(组1，得分0或1，$N=61$)，这组中“超过”题项梯度(01)的研究对象所占的比例为$(2+1)/61=0.05$。“超过”题项梯度(12)的研究对象所占的比例为$1/61=0.02$。对题项梯度(01)来说，所有的比例从最低到最高余分分组是逐渐变大的。对题项梯度(12)来说，在第一和第二余分分组间存在值为0.02的违反(低于默认值0.03)。注意，对于题项梯度(01)而言，可以对7个余分分组构成的所有21个配对组进行比较。对于题项梯度(12)而言，第二余分分组正向回答的比例为0.00，因此该组不参与与其他组的比较，除了一个例外。这样就剩下了由6个余分分组构成的15个配对组比

较。前面提到的例外是第一和第二余分分组的比较,对于这个题项梯度函数适合的比较总数是 16 个,并且适合比较的总数为 21+16=37。

表 8.15 对"罢工"题项单调同质性的检验结果

Strike · Joining unofficial strikes

Restscore Group	Low		High	N	Frequencies per item value 0	1	2	Mean	Proportions of positive responses per item step ≥1	≥2
1	0	—	1	61	58	2	1	0.07	0.05	0.02
2			2	126	108	18	0	0.14	0.14	0.00
3			3	182	129	48	5	0.32	0.29	0.03
4			4	323	212	106	5	0.36	0.34	0.02
5			5	206	70	126	10	0.71	0.66	0.05
6			6	131	42	70	19	0.82	0.68	0.15
7	7	—	8	81	6	52	23	1.21	0.93	0.28

使用 50 个研究对象以上的余分分组对 DM 进行检验,发现存在若干数量的模型违反,并且有一些在统计上是显著的。结果如表 8.16 所示。

表 8.16 使用至少 50 个研究对象的余分分组进行的双重单调性检验结果(IRF 不相交),最低违反界限设为 0.03

Summary per item for check of nonintersection via restscore groups Minimum violation = 0.03 sig. level = 0.05 Minimum group size = 50

	Item H	# ac	# vi	# vi/# ac	maxvi	sum	sum/# ac	zmax	# zsig	crit
Petition	0.41	72	2	0.03	0.09	0.17	0.002 4	3.08	2	46
Boycott	0.44	64	5	0.08	0.09	0.33	0.005 1	2.70	1	48
Demons	0.46	64	5	0.08	0.09	0.35	0.005 5	3.08	1	49*
Strike	0.49	72	1	0.01	0.07	0.07	0.001 0	1.21	0	9
Occupy	0.35	72	3	0.04	0.07	0.17	0.002 4	1.22	0	23

最大的违反(maxvi)存在于题项梯度"游行"(01)和"请

愿"(12)之间(表 8.17),是在第四余分分组中(余分为 3):违反值为 0.09(0.96－0.86),z 值为 3.08 并且在统计上显著。然而由于 crit 值还是相当低的,所以并没有足够的理由来舍弃任何一个题项。

表 8.17　最大的模型违反的详细信息

I = (Demons>1) with list below

J = (Petition>2)

Restscore			Frequencies for				vi	z	Proportions of positive responses per item step	
	Low/High	*N*	item	step	pair	IJ				
Group			00	01	10	11			Demons>1	Petition>2
1	0—0	194	51	70	25	48			0.38	0.61
2	1—1	314	21	63	37	193			0.73	0.82
3	2—2	268	9	53	25	181			0.77	0.87
4	3—3	200	2	7	25	166	0.09	3.08	0.96	0.86
5	4—6	134	2	3	5	124			0.96	0.95
		-----	--	---	---	---	----			
Total		1 110	85	196	117	712	0.09		0.75	0.82

使用 P(+,+)和 P(－,－)矩阵进行的 DM 检验没有找到任何严重的模型违反的迹象。实际上,P 矩阵根本没有显示出模型违反。根据莫坎的方法计算的信度是 0.72。

第 6 节 | 政治行为量表

表 8.18 给出了 1 110 个没有缺失值的研究对象在政治行为量表上的量表值。大部分研究对象的量表值在 4 到 8 之间:他们签署请愿,也可能参加联合抵制和合法游行,但是他们对于参加非正式罢工和占领建筑物和工厂的行动不是那么热心。量表得分的分布接近于正态分布(偏度 = −0.16,峰度 = −0.22)。很少的人对所有五个题项给出同样的答案:只有 7 个研究对象任何行动都不愿意参与,只有 20 个研究对象所有的行动都参与过。

表 8.18 非议会政治行为量表的量表得分以及哥特曼误差数

	量表得分		哥特曼误差	
	频次	百分比	频次	百分比
0	—	—	552	50
1	—	—	147	13
2	—	—	162	15
3	—	—	105	9
4	—	—	71	6
5	7	1	37	3
6	25	2	9	1
7	64	6	5	0

续表

	量表得分		哥特曼误差	
	频次	百分比	频次	百分比
8	84	8	5	0
9	173	16	11	1
10	181	16	5	0
11	260	23	1	0
12	148	13	—	—
13	110	10	—	—
14	38	3	—	—
15	20	2	—	—
均值	10.4	100%	1.4	100%
标准差	2.0		1.9	

一半的研究对象不存在哥特曼量表的误差，只有一个人(第34号研究对象)制造了11个误差：他/她参与过占领活动，也可能签署请愿书，但是绝对不会参与联合抵制和合法游行，或者参与非正式罢工(表8.19)。

表8.19 第34号研究对象的回答模式，包括了11次模型违反，根据10个定序排列的题项梯度进行了二分编码

P_{01}	P_{12}	B_{01}	D_{01}	S_{01}	O_{01}	B_{12}	D_{12}	S_{12}	O_{12}
1	0	0	0	0	1	0	0	0	1

现在可以来确定谁的政治态度更加积极。我们可以把量表值和哥特曼误差联系起来。没有理由去期待这两个变量会有什么高度相关，它们也只是具有稍微的相关度：$r=0.085$($p=0.005$, $N=1\,110$)。较高的量表值意味着具有较高的概率发生哥特曼误差。

第9章

余 论

在本章中我们讨论两个余论。第一个是比较我们的定序 IRT 模型与其他测量模型,第二个是如何处理不具有区分度的回答模式。涉及其他的测量模型时,我们一方面会讨论到信度分析和因子分析,另一方面会讨论参数 IRT 模型,特别是单参数 logistic 模型,也被称为拉希模型(Andrich, 1988; Ostini & Nering, 2006; Rasch, 1960/1980)。这些其他的模型可能会得出不同的结果,因此我们需要理解背后的原因。不具有区分度的回答模式在这些其他模型中的作用与在 IRT 模型中不同,所以我们需要讨论在 IRT 模型中对这种回答模式的处理。随后我们讨论两个实际问题:我们如何使用研究对象的量表值?以及我们可以使用回答类别的数量不同的题项吗?我们以一些最后评论结尾。

第1节 | 信度分析

同质性比上信度

定序累计量表分析中的同质性概念不同于信度分析中的信度概念。同质性是指:不同的题项测量的是同一个潜在特质。信度是指:当再次询问同样的问题时,研究者仍能得到同样的回答。因为再次询问同样的问题不太好——至少不要在相同的调查中询问——想要建立测量的信度是比较困难的。因此研究者会使用相似的(平行的)问题,把它们看作是相同的问题。这样可以建立信度,并被理解为同质性。

莫坎信度估计

莫坎(Mokken, 1971)提出了一个创造性的对于累计量表的信度估计。他使用 P(+,+)矩阵——我们曾在第6章里用它来检验 DM 的模型假设——来估计对特定题项给出两次相同答案的研究对象比例。这个比例很明显是算不出来的,但是 P(+,+)矩阵的对角线可能会提供一些信息。

如果 P(+，+)矩阵符合双重单调性的假设，那么每行或者每列的单元格比例是单调递增的，那么对角线上的比例的值可以用来做外推估计(对于外围的两个题项而言)，或者内推(对于其余的题项而言)。这些比例可以用来计算信度(即真实的方差与实际观测到的方差的比值)。附录第 3 部分解释了莫坎的这个方法，并且用第 5 章美国人宗教信仰的例子中的五个二分题项作了一个计算说明。

相关系数及其问题

假设所有的题项具有相同的标准差，那么信度就可以很容易地用所有题项的平均相关系数，r_{av}来计算：

$$\text{信度}\ \rho = \frac{k \times r_{av}}{1 + (k-1) \times r_{av}} \quad \text{(这里 } k \text{ 是题项数量)}$$

题项之间的平均相关系数越高，信度也越高。也可以通过往量表中加入题项来提高信度，只要平均相关系数不会下降太大。增加题项通常不会增加量表的同质性系数。使用相关系数矩阵意味着研究者假设对他的题项的回答是沿着一个定距量表来测量的。我们稍后讨论这一假设的合理性。

很早以前就有人提出(Carroll，1945；Ferguson，1941)，当题项不是互相平行的时候，即题项的均值和标准差不相等的时候，使用相关系数就会存在问题。我们用一个假想的数据集来阐述这个问题，如表 9.1 所示(另一个相似的例子参见 van Schuur，2003)。一百名研究对象回答六个二分题项。7 个理想的累计回答模式的频次在最后一列给出。假设 D，

表 9.1 构成一个理想累计量表的假想数据集

回答类型	*A*	*B*	*C*	*D*	*E*	*F*	回答模式的发生频次
1	0	0	0	0	0	0	1
2	0	0	0	0	0	1	4
3	0	0	0	0	1	1	6
4	0	0	0	1	1	1	34
5	0	0	1	1	1	1	4
6	0	1	1	1	1	1	6
7	1	1	1	1	1	1	45
	45	51	55	89	95	99	100

E 和 F 分别是三个变量，测量在几乎每个人都参与的三次选举中的投票行为，而 A，B 和 C 三个变量测量三次复选的投票行为。后三个题项的普及度很高——大部分研究对象都参加了这些活动（$p_D=0.89$，$p_E=0.95$，$p_F=0.99$）——而头三个题项的普及度则比较低（$p_A=0.45$，$p_B=0.51$，$p_C=0.55$）。表 9.2 列出了相关系数和同质性系数。所有题项对的同质性都是完美的，但是相关系数的大小则不同，从对角线左右的最大值到左下角的最小值。根据这些题项的边缘频次分布，这些相关系数已经是可能的最大值了。

在这个数据集中，六个题项构成的量表的信度（克朗巴赫系数，Cronbach's α）是 0.82，但是如果删除题项 6 以后，信度会提高到 0.85，或者事实上，删除最后三道题项后的信度会上升到 0.96。因此，最后三个题目拉低了前三道题目的信度，虽然所有的六个题项构成了一个理想的累计量表。

表 9.2 表 9.1 中数据的相关系数和同质性系数

	相关系数						同质性系数				
	A	B	C	D	E		A	B	C	D	E
B	0.89					B	1.0				
C	0.82	0.92				C	1.0	1.0			
D	0.32	0.36	0.39			D	1.0	1.0	1.0		
E	0.21	0.23	0.23	0.65		E	1.0	1.0	1.0	1.0	
F	0.09	0.10	0.10	0.29	0.44	F	1.0	1.0	1.0	1.0	1.0

第 2 节 | 因子分析

因子分析是第二个应该与我们的定序 IRT 模型进行比较的模型。[20]在更宽泛的意义上说,因子分析可以看作是信度分析的一般化。在一组题项中,因子分析试图寻找一些题项的聚类,这些题项中的任何一个都可以构成一个可信量表的组成部分,这些题项具有较高的类别内(或因子内)的相关系数。使用相关系数存在的问题在因子分析中与在信度分析中同样明显,因为相关系数的值取决于变量的分布。对表 9.1的数据进行因子分析得到了两个特征值大于 1 的因子,而不是一个。表 9.3 显示了最大方差旋转法得到的结果,前三个难度较高的题项与后三个较容易的题项分别构成了不同的因子。

表 9.3 表 9.1 中数据集经最大方差旋转后得到的因子负载

	因子 1	因子 2
V1	0.93	0.10
V2	0.97	0.13
V3	0.94	0.16
V4	0.31	0.76
V5	0.13	0.88
V6	−0.03	0.72

无论是关于这个数据集的信度分析结果还是因子分析结果,都严重挑战了这六个变量的单一维度性。赞同因子分析的人可能会在数据中找到一个单一维度累计模式,因为这六个题项在第一个未旋转过的因子上的负载都很高(此处未展示),而且这两个旋转后得到的因子可能被理解为难度高的因子,理解一个因子的最低界限的特征值太低,仅为1.00,或者是应该使用多项相关而不是积矩相关。由此我们可以理解:对适合累计模型的数据进行因子分析并不是那么直接的。

因子分析和信度分析都假设所有的题项都是平行的,即,具有相同的频次分布(相同的均值和标准差),但是这一假设在累计量表中基本上不会满足:题项的频次分布确实存在差异。这种差异就是累计量表形成的原因,也是为什么对二分类数据的因子分析很难解释的原因。题项的难度排序通常具有重要的理论解释,但是信度和因子分析中并没有考虑到这一点。比如,在第2章测量富裕程度的量表中,如果拥有CD播放机比拥有洗碗机更容易,那么研究者可以推断消费者购买CD播放机要早于购买洗碗机。这样的推断是不可能从信度分析或者因子分析中得出的。如果题项确实构成了一个累计量表,或者研究者预期这些题项可以构成累计量表,那么在模型中考虑到前面讲的这一点是有必要的。

定序IRT模型相比因子分析的第二个优势在于因子分析总是有一个解答,即研究者总是可以发现至少一个因子,除非所有的题项彼此完全不相关。因为总是会存在至少一个特征值最大的因子,特征值大于1,并且这个因子并不是那么容易解释。然而在定序IRT分析中,所有的**H**(ij)值都可能低于用户自定义的最低界限,这种情况下是不存在量表的。

第3节 | 参数IRT模型:拉希模型

定序IRT模型除了应该与信度分析和因子分析相比较，还应该与参数IRT模型进行比较,因为后者在一些学科领域比如教育学中的应用更加广泛。本书讨论的IRT模型属于非参数模型,因为它们的参数仅仅是题项的排序和研究对象的排序。相反,在参数IRT模型中,研究对象和题项的参数是在一个定距量表上的取值。但更重要的是,一种特别的参数模型——简单logistic模型或者单参数logistic模型——具有特定的测量特征。在这个模型中,对研究对象的测量不依赖于使用特定的题项,并且该模型允许一些量表的比较,这些量表包括了一些相同的题项,但不是全部题项都相同。这种模型有时被称为拉希模型,是根据它的创立者,丹麦统计学家格奥尔格·拉希来命名(Georg Rasch, 1960/1980)。

拉希模型最优越的特性叫作特定客观性或者题项(研究对象)不变性。这就是说,题项参数的估计不会随着特定研究对象的改变而改变,而且研究对象参数的估计也不会随着特定题项的改变而改变。特定客观性的原因是由于将IRF作为平行logistic方程来估计。这一特性有若干优势。比如,研究者可以截取某些题项进行检验,因为研究对象不需要回答所有的题项,只需要回答具有大致相同的量表值的题

项即可,然后把这些值作为研究对象的取值。这样就可以对不同的检验进行合并,只要存在一些共同的题项来做调整。特定客观性的特性和可以做截取性检验这两点使得拉希模型在心理学和教育学测量中成为一个重要的模型,这些领域要求相应的测量必须准确以及具有足够的信度,以便对单独研究对象进行讨论。因为对准确度和信度的要求,基于拉希模型的检验通常比大多数调查中的测量工具要长些。

如果拉希模型拟合数据,那么它将是一个非常理想的测量模型。但是模型拟合只有当假设成立的条件下才能达到,而平行 logistic IRF 的假设是非常(经常过于)严格的。拉希模型可以被看做是 DM 模型的一个特例,但是也有一些保留。首先,拉希模型不使用不具区分度的研究对象,即,对所有题项给出相同回答的研究对象。这点稍后会详细讨论。第二,即使数据没有什么结构(即随机数据),拉希模型也可以拟合这些数据(Wood, 1978)。但是随机数据不符合定序 IRT 模型,因为同质性不够。另一方面,一个理想的哥特曼量表不是一个好的拉希量表,因为当不具区分度的研究对象和题项被剔除后,基本上就没有剩下什么内容了。

第4节 对美国人宗教信念数据应用其他测量模型

信度分析和因子分析都需要用到相关系数矩阵，如表9.4a所示。为了比较，表9.4b给出了同质性系数矩阵。同质性系数的值高于相关系数，并且在相关系数矩阵中，行和列的题项是按照普及度来排序的。不同题项之间在分布上的差异越大，它们之间的相关系数越小。

表9.4 相关系数和同质性系数，所有研究对象，$N=1\ 200$

	相关系数					同质性系数			
	地狱	重生	天堂	灵魂		地狱	重生	天堂	灵魂
重生	0.32				重生	0.36			
天堂	0.64	0.39			天堂	0.96	0.52		
灵魂	0.34	0.35	0.43		灵魂	0.86	0.80	0.72	
上帝	0.34	0.27	0.49	0.39	上帝	0.87	0.63	0.86	0.40

a. b.

该数据的信度分析显示，克朗巴赫系数为0.743。如果把最差的题项（重生）去掉，克朗巴赫系数为0.738。但是同一数据的因子分析只能得到一个因子，并且所有的五个因子负载都很高（见表9.5）。

表 9.5 第一个因子的因子负载

地 狱	0.75
重 生	0.62
天 堂	0.85
灵 魂	0.69
上 帝	0.69

然而,美国人宗教信念的题项不适合用拉希模型来分析。[21]

第5节 | 不具区分度的回答模式

虽然题项构成了一个很好的累计量表，一个可信的检验及一个很好的因子，但是为什么拉希模型和数据拟合得就不好呢？原因在于：28个对任何题项都不正向回答的研究对象以及716个对任何题项都正向回答的研究对象的信息是否被使用(见第5章表5.8)。在拉希模型中，人们需要在题项之间做出区分来确定题项间的距离，同样地，题项也需要区分度(即，每个研究对象都给出正向回答或者负向回答的题项是无用的)来确定研究对象之间的距离。因此拉希模型只使用了456个具有区分度的研究对象，并且发现了一个拟合很差的量表。

对于信度分析来说，没有比研究对象每次都给出相同(可靠)的答案更好的了。没有区分度的研究对象构成了信度分析的支柱。只有当研究对象在多个题项上给出相同的特定的分数才能得到高的信度系数。因此，当然没有理由来删除这些研究对象。然而，如果我们只对456个具有区分度的研究对象重复信度分析，我们会得到一个0.19的信度系数，并且只有两道题项——“地狱”和“天堂”——可以构成一个信度为0.60的量表。因子分析则得到三个特征值大于1的因子。定序的IRT模型不适用于这些研究对象。表9.6

给出了这个分析中的相关系数和同质性系数。

表 9.6 相关系数和同质性系数,具有区分度的研究对象,$N = 456$

	相关系数					同质性系数			
	地狱	重生	天堂	灵魂		地狱	重生	天堂	灵魂
重生	−0.50				重生	−0.64			
天堂	0.43	−0.01			天堂	0.90	−0.01		
灵魂	0.10	0.11	0.18		灵魂	0.48	0.40	0.40	
上帝	0.10	−0.04	0.30	−0.07	上帝	0.51	−0.15	0.69	−0.08

a. b.

"地狱""天堂"和"上帝"三个题项可以构成一个累计量表,另外两个题项("重生"和"灵魂")则构成第二个量表。注意,"灵魂"题项不属于第一个量表,因为六个题项对之一("灵魂","上帝")的同质性系数是负的(−0.08)。

就像这个例子显示的,在分析中是否纳入不具有区分度的研究对象会造成完全不同的结果。如何处理不具有区分度的研究对象引起了很多严肃的讨论:他们是否可以证明信度模型的合理性?或者他们是否人为地提高了量表的同质性系数?因为不具区分度的研究对象不能够推翻数据中假设的累积性。对于题项数量很少的量表而言,删掉不具有区分度的研究对象看起来像是一个不好的做法,因为他们通常占据了研究对象总数的相当比例。但是对于具有很多题项的量表而言,没有区分度的研究对象只占研究对象总体的较小比例,即使删除了他们,累计模型依然适用。除了建议研究者要谨慎地决定是否在分析中纳入或排除不具区分度的研究对象之外,给出明确的处理建议是不可能的。

第6节 | 一些实际问题

量表得分:定序或者数值型?

在本书中我已经强调过,我们的分析方法都是关于定序的而不是数值型的,结果是定序的题项或者题项梯度,以及(部分地)定序的研究对象(得分相同的研究对象是无法区分的)。但是很多统计方法,比如回归分析,要求必须在数值型的量表上测量研究对象。我们可以用这些量表得分——我们量表分析的最终结果——进行下一步的统计分析吗?

基于我们对从伦西斯·里克特1932年具有开创性的文章以来文献的研究,我认为上述问题的答案是肯定的。里克特将 −2, −1, 0, +1, +2 这些回答分类视为五分自评量表中删截的整数(可以在这些值上再加数字,比如2或3,不会改变这些分数的数值含义),然后比较 z 值。这样会导致原来 z 值的定序转换。他得出结论说,基于删截数据的统计结果不比原来的结果差。他的这一发现是20世纪定量调查研究的重要成就,因为五分自评量表现在被当做定距数据来使用。在使用定序 IRT 时,我们就沿用里克特的这一发现。信度分析和定序 IRT 都用所有题项的未加权总分作为研究对

象的量表得分。如果定序转换不会严重改变对之后统计分析结果的理解,那么我们也可以将里克特的发现运用到扩展研究对象量表值上,把这些数值作为定距数值并且运用到最常见的线性统计模型中去,比如因子分析和回归分析。

具有不同回答类别的题项的运用

在本书中,正如其他大多数介绍类的读物中一样,构成量表的所有题项都具有相同数量的类别。但是这一点是必要的吗? 对此还没有明确的说法。前面讲过,一个累计量表的基本构成要件是递增的基准值或者题项梯度。这意味着在一个量表中使用具有不同类别的题项在方法上应该是可行的,这点在技术上已经实现了。注意,模型对于不同题项答案类别之间的距离排序上并没有什么前提假设。

第7节 一些最后评论

定序IRT起始于莫坎1971年发表的论文,之后就被称为莫坎量表分析。此外,莫坎还发表了若干篇论文(如,Mokken, 1997; Mokken & Lewis, 1982),包括他的一些学生发表的论文(如,Niemöller & van Schuur, 1983; van Schuur, 2003)。然而,大多数新的发展来自他的朋友及其同事伊沃·穆伦纳尔(Ivo Molenaar),穆伦纳尔的学生克拉斯·赛斯玛(Klaas Sijtsma),以及赛斯玛的学生如巴斯·赫姆克尔(Bas Hemker)和安德里斯·范德·阿尔克(Andries van der Ark)。这些方法仍然处于发展之中。赛斯玛和穆伦纳尔的书(Sijtsma & Molenaar, 2002)是目前对这一方法最标准的介绍。

莫坎量表分析的统计软件首先是由翰·霍尔(Han Hol)发展出来的(在20世纪60年代),叫做Scale Analysis Method Mokken(SCAMMO),之后(20世纪70年代)由维姆·范·霍博肯(Wim van Hoboken)和皮埃尔·德贝(Pierre Debets)纳入SPSS的Statistical Appendix(STAP)中,名称为MOKKEN SCALE。在罗布·莫坎、伊沃·穆伦纳尔、克拉斯·赛斯玛和韦杰勃朗·凡舒尔的监督下,彼得·博尔(Peter Boer)发展出了一个新的windows程序,MSP。其他

人发展出了在其他语言环境下使用的程序，或者是 SPSSX，STATA，或者 R 的改编版本：金马和泰尔鲁姆（Kingma & Taerum，1989），里瓦斯和马丁内斯（Rivas & Martinez，1992），里瓦斯、马丁内斯和伊达尔戈（Rivas，Martinez & Hidalgo，1996），范德·阿尔克（Van der Ark，2007）以及维西（Weesie，1999）。

最早的应用只限于一小部分研究者，包括莫坎、穆伦纳尔和赛斯玛，他们获得了来自下面这些关于量表分析和维度分析的暑期课的帮助：在安娜堡举办的各大学政治和社会研究联合暑期学校（Inter-University Consortium for Political and Social Research Summer School），艾塞克斯社会科学数据分析和收集暑期学校（the Essex Summer School in Social Science Data Analysis and Collection）以及布鲁塞尔社会科学定量分析计划（the Brussels Quantitative Analysis in the Social Sciences program）。现如今的模型应用非常广泛，尤其是在医学社会学领域。在经济心理学、政治科学和教育学中也有应用。可以期待定序 IRT 方法会被编入社会科学研究方法的教材，也会被纳入主要的统计软件之中。

附 录

H 系数的零分布

某个规模为 N 的样本具有特定数值的同质性系数(如 0.50),但是该样本可能是从 **H** 系数为 0.00 的总体中抽取的,这种情况发生的概率是可以被估计的。为了估计这个概率,我们需要知道当所有的回答在统计独立的情况下,**H** 系数的分布。在一个具有固定边缘分布的 2×2 的交互表中,我们可以根据误差单元格和(1, 1)单元格来计算 **H**(ij)系数:

$$\mathbf{H}(ij)=\frac{p(ij)-p(i)*p(j)}{p(i)\times[1-p(j)]},\text{或者}\frac{\Delta(ij)}{p(i)\times[1-p(j)]}$$

莫坎计算出了 **H** 分布:均值为 0.00,方差则取决于题项的难度,可以写成下面这一公式(Mokken, 1971:162):

$$\sigma^2=p(i)[1-p(i)]p(j)[1-p(j)]$$

因此这个分布的标准差 σ 就是它的平方根。

莫坎定义 $\Delta(ij)^*=\sqrt{(n-1)}\times\Delta(ij)/\sigma$,对于足够大的样本规模 N,就是标准正态分布。在之后的文章中,$\Delta(ij)^*$ 被重新定义为 $Z(ij)$。因此如果 $Z(ij)$值大于 1.64,那么根据标准正态分布(单侧检验),观测到的同质性系数在整体中大于 0.00 的概率就超过 5%。我们用表 5.2(第 5 章)来展示如何计算 $Z(ij)$:

表 A.1 “重生”(横向),“地狱”(纵向)

	0	1	总数	
0	157	134	291	24.25%
1	185	724	909	75.75%
总数	342	858	1 200	
	28.5%	71.5%		

$p(ij)=724/1\,200=60.33\%\quad p(i)=0.757\,5\quad p(j)=0.715$

$$p(i)\times p(j)=0.541\,6$$

$\Delta(ij)=p(ij)-[p(i)\times p(j)]=0.603\,3-0.541\,6=0.061\,7$

$\sigma^2=(0.242\,5*0.757\,5*0.285*0.715)=0.037\,4$,因此$\sigma^2=0.193\,4$

$$\sqrt{1\,200-1}=\sqrt{1\,199}=34.627;\ 34.627*0.061\,7=2.136;$$

$$2.136/0.193\,4=11.05$$

因此$\Delta(ij)^*$或者$Z(ij)=11.05$。这个值远大于3,所以**H**(ij)系数是非常显著的。

注意,莫坎比较的是(1, 1)或者(+, +)单元格的观测值和期望值,而不是比较误差单元格的观测值和期望值。但是一个具有给定边缘分布的2×2交互表只有一个自由度,因此比较误差单元格或者(1, 1)单元格是没有太大分别的。

整个量表的Z值可以这样来计算:

$$Z=\frac{\sqrt{N-1}\sum_{i=1}^{k-1}\sum_{j=i+1}^{k}\Delta(ij)}{\sqrt{\left(\sum_{i=1}^{k-1}\sum_{j=i+1}^{k}\sigma ij^2\right)}}$$

对于单个题项**Z**(i)值的计算方法也是一样的,只对包括题项i的Δ和σ加总$\Delta(ij)$和$\sigma(ij)$。

防止随机结果的出现

当新的题项不断地加入量表时,我们想把选择的标准变得更加严格。这一点是可以实现的,只要不断地提高统计显著性的标准。起初,每个题项的 $\mathbf{H}(i)$ 值都是显著大于 0 的。我们使用 $\mathbf{H}=0$ 的总体中 $\mathbf{H}$ 系数标准差的 z 值。之后使用使用者自定义的显著度水平作为它的补充,而不是 α 水平,一般定为 5%。当量表规模逐渐变大时,我们想要把这一概率,比如 5%,逐渐降低到一个更低的值(α^*),以提高置信度水平。莫坎提出了下面这些步骤:

1. 选择一个通常的置信水平 $1-\alpha$(比如,$\alpha=0.05$)。

2. 使用另外一个 α 水平(α^*),将使用者自定义水平除以题项对数量[$\alpha^*=\alpha/\frac{1}{2}(n-1)n$]。

3. 每往量表中加入一个题项,就把分母的值提高 $k(r_j)$,r_j 是第 j 步余下的题项数目,需要从中选出下一个加入量表的题项。开始的题项集起始于第 1 步,从第 2 步开始加入第一个题项。对选择第 $(j-1)$ 个题项的检验,此时已经过了 h 步,是用下面这个置信水平来计算的:

$$\alpha^* = \frac{\alpha}{\frac{1}{2}(n-1)n + \sum_{j=2}^{h} r_j}$$

让我们用一个例子来说明。假设有 10 个题项,彼此之间正相关并且可能构成一个累计量表的一部分,使用者选择的 alpha 水平是 0.05。这 10 个题项可以构成 45 个题项对,因此评价第一对题项的 α^* 水平为 0.05/45 =0.001 11。选完第一个题项对之后,就从余下的 8 个题项中选择第三个题项。那么往现有的两个题项构成的量表中加入第三个题项,相应的 α^* 水平是 0.05/(45+8)=0.000 94。选择第四个题项的 α^* 为 0.05/(45+8+7)=0.000 83,以此类推。在许多实际应用中,当样本规模不是太小(如大于 100),并且题项的难度不是极端值(比如处于 0.1 和 0.9 之间),这里介绍的做法不会影响结果。

计算信度

两次给出相同答案的概率,正如测试—再测信度所要求的,通常不能够通过询问两次同样的问题来得到。莫坎建议使用P(+,+)矩阵来计算这一概率,这个矩阵给出了研究对象在题项对上给出正向回答信息的比例。如果题项符合双重单调性的要求,当题项按照难度来排序时,P(+,+)矩阵中正向回答的概率将从左到右,从上到下递增。对角线上的单元格可以被理解为对相同题项给出两次相同正向回答的研究对象比例($p(ii)$)。这些比例可以用外推法(对于第一个和最后一个题项)或者内推法(对于中间的题项)。让我们用美国人的宗教信念做例子。P(+,+)矩阵的对角线上的值给出了外推或者内推的值(用星号标出)。

表 A.2 P(+,+)矩阵,++概率及信度估计值

题　项	P	A=“地狱”	B=“重生”	C=“天堂”	D=“灵魂”	E=“上帝”
		0.71	0.76	0.85	0.94	0.94
A=“地狱”	0.71	0.58*	0.60	0.71	0.71	0.71
B=“重生”	0.76	0.60	0.64*	0.70	0.75	0.74
C=“天堂”	0.85	0.71	0.70	0.78*	0.83	0.84
D=“灵魂”	0.94	0.71	0.75	0.83	0.91*	0.91
E=“上帝”	0.94	0.71	0.74	0.84	0.91	0.91*

对题项 E 进行外推法估计需要建立在如下假设之上：对于那些在题项 E 之前的题项，它们的难度之间的关系（$p(i)$ 值），和由它们构成的题项对的正向回答的比例之间的关系是相同的。即，对于 p(EE)的外推：

$$\frac{p(\mathrm{E})-p(\mathrm{C})}{p(\mathrm{D})-p(\mathrm{C})}=\frac{p(\mathrm{EE})-p(\mathrm{CE})}{p(\mathrm{DE})-p(\mathrm{CE})}$$

从上式中我们得到 $(0.94-0.85)/(0.94-0.85)=[p(\mathrm{EE})-0.84]/(0.91-0.84)$，或者 $p(\mathrm{EE})=0.91$。

对于 p(BB)的内推：

$$\frac{p(\mathrm{C})-p(\mathrm{B})}{p(\mathrm{C})-p(\mathrm{A})}=\frac{p(\mathrm{BC})-p(\mathrm{BB})}{p(\mathrm{BC})-p(\mathrm{AB})}$$

从上式中我们得到 $(0.85-0.76)/(0.85-0.71)=[0.70-p(\mathrm{BB})]/(0.70-0.60)$，或者 $p(\mathrm{BB})=0.64$。

对这些值的估计现在被用于信度的公式中，由莫坎提出(Mokken, 1971:145，方程 2.28)。这个公式中的分子是对于真实分数的方差估计，分母则是对于观测分数的方差估计。赛斯玛和穆伦纳尔(Sijtsma & Molenaar, 1987)指出，与其他计算信度的方法相比，这个方法更具优势。

$$\text{信度}=\frac{\sum_{i=1}^{k}[p(ii)-p(i)^2]+2\sum_{i=1}^{k-1}\sum_{j=i+1}^{k}[p(ij)-p(i)^{*}p(j)]}{\sum_{i=1}^{k}\{p(i)[1-p(i)]\}+2\sum_{i=1}^{k-1}\sum_{j=i+1}^{k}[p(ij)-p(i)^{*}p(j)]}$$

与其他大多数信度分析一样，信度随着(同质性)题项数量的增加而增加。一个很短的量表(较少的题项)的同质性系数可以很高，但是信度系数可以很低。如果在调查中一组题项的信度高于 0.60，那么用这些题项构成量表就是可以接受的。

注释

[1] 临近回答可能看起来没有用处,因为负向回答的不确定性。但是在一些情形下,比如在分析偏好、儿童发展,或者政治的情况下,询问临近回答性质的问题是有用处的。在那些案例中,负向回答包括两种相反意思中的一种。举例来说,"你喜欢喝加了一块方糖的咖啡吗?"("不,我喜欢喝不加糖的咖啡",或者"不,我喜欢喝加了更多糖的咖啡")。"宝宝会爬了吗?"("不,宝宝还不会爬",或者"不,宝宝已经不再爬了")。"投票是人们在政府中拥有话语权的唯一途径吗?"("不,投票不是一种途径",或者"不,还有更多的途径")。这类数据不能用这本书介绍的模型来分析。然而它们可以用一种叫作展开模型的 IRT 测量模型来分析,这种模型是由库姆创造的(参见 Coombs, 1950, 1964;以及一些他的继承者,比如 Andrich & Luo, 1993; Roberts, Donoghue, & Laughlin, 2000)。定序展开模型,类似于本书介绍的模型,是由凡舒尔(van Schuur, 1993)和凡舒尔与基尔斯(van Schuur & Kiers, 1994)发展出来的。循环展开模型,也被称为环状模型(如 Plutchik & Conte, 1997),已经由布朗(Browne, 1992)和其他人发展出来。定序循环展开模型也已经发展出来了(Mokken, van Schuur, & Leeferink, 2001)。

[2] 欧洲和世界价值观调查(European and World Values Study)询问了很多此类的问题。

[3] 可以询问 8 个有关人们是否拥有某项设备或者进行某项活动的问题(即,如果他们对问题给出了正面回答,那么他们对于较容易的题目同时也会给出正面回答)(Sanders & van Schuur, 1998)。

[4] 累计量表或者蕴含量表的术语使用根据不同的学科而有所不同。

[5] 哥特曼相信具有两个答案或者选项的变量(比如,是/否,同意/不同意)比具有五个选项的变量要易于回答。有些题项本质上就是二分的,比如具有正确和不正确选项的问题。但是在调查中经常通过合并三个或更多选项把题项变成二分类的。例如,如果问题(常被称为题项)具有五个答案选项,1=非常同意,2=同意,3=不确定,4=不同意,5=完全不同意,1 和 2 可以被编码为 1=同意,3、4 和 5 可以被编码为 0=不同意,这样题项就变成了二分类的。回答 1 为正向回答,回答 0 则为负向回答。

[6] 只有我们确立了题项的难度次序,确定模型违反的数量才是可能的。以前用更改回答以得到更完善的回答模式来定义模型违反数量的做法意味着这些更改可能影响到题项间的难度次序。然而,在使用一个研

究对象和两个题项间的传递关系来定义模型违反的情况下，上述问题就不存在了。在这种模型的实际应用中，研究者既可以利用提前确立的项目的理论次序，也可以尝试可能存在的不同的难度排序。在本书中我们不会更多地涉及这一话题，而是假设项目的难度次序是已知的。

[7] Err(exp)是根据 $[1-p(i)]\times p(i)\times N$ 来计算的，其中，$p(i)$和 $p(j)$分别是对应题项正向回答的相对频率，N 是数据集的样本规模。**H** 与古德曼和克鲁斯卡尔(Goodman & Kruskal, 1979)的系数 λ(lambda)具有同样的意义，即误差测量的比例缩减。它也可以理解为在控制住两个变量的边缘分布的条件下，两个变量间的相关系数与可能的最大相关系数的比值。

[8] 如同莫坎指出的，研究对象表现出来的量表分值与他们真正的潜在量表分值是高度相关的(Mokken, 1971:140—141)。

[9] 我们可以用因子分析来类比理解这种方法。第一个非旋转因子被视为测量单一潜变量的对所有题项的最好结合，但是经旋转后，我们可能使用同样的题项来测量所研究的潜变量的特定方面。

[10] 该数据集可以从主要的数据资料库得到，比如(美国)Inter-University Consortium for Political and Social Research，或者(德国)GESIS-Leibnitz-Institut für Sozialwissenschaften. 这样读者就可以学习本书的例子并且可以手动检查所有的计算。

[11] 我们在本书中不涉及缺失数据的问题，除了规定缺失数据可以被视为“不属于正向回答”，因此是一种负向回答。或者具有任何缺失数据的回答者可能会被删除。关于莫坎量表中缺失数据填补问题的更多信息，请参阅于斯曼(Huisman, 1998, 2000)。

[12] 前人研究把 IRF 称为“题项特征曲线”或者“痕迹线”。

[13] 出于历史兴趣，莫坎提出了一个检验来考察两个样本的同质性系数是否相同(Mokken, 1971)。他提出在一组研究对象中观察是否潜在特性的某些指标随着时间变化——如社会化过程——而表现出更高的同质性是非常有趣的。然而对同质性差异还有另外一种理解：研究对象的量表值的方差越小，他们的回答模式就越符合局部随机独立性假设。方差的差异可以导致同质性的差异。这些不同的理解是很难区分彼此的，因此这一检验在实际中也不再使用。

[14] 也存在具有三个或更多个分类、但不是定序的题项，如“你相信上帝吗？(0)不相信，(1)不知道，(2)相信。”这里，“不知道”是否处于其他两个回答类别之间是不清楚的，它只是简单地不同于其他两类。像这种题项在 IRT 的情境下不被称为多分类问题，也不在本书的讨论范围之内。

[15] 这种类型的题项也叫做五分自评量表，或者里克特量表(Likert Scale)。

"量表"这一术语来自里克特的看法,五个类别可以看做一个定距量表。在本书中我们所说的"量表"术语只是指问题的汇总,类似心理学家使用"测试"这一术语。

[16] 如果想得到这些信息,直接询问一个关于身高的五分类的问题比询问两道三分类的问题更加直接。我们并不能提前知道两个或多个题项测量的是否是同一个潜变量。这里对身高的处理是说教的,只是为了说明在一个潜在的量表里如何合并多分类的题项。

[17] 假设读者可以自行获得数据,并且可以用统计软件包做交互表。

[18] 通常来说,研究者应该调整回答类别的编码顺序,与要测量的潜变量的方向保持一致。

[19] 比如,一道五分类的里克特量表(0, 1, 2, 3, 4)具有四个题项梯度:(01), (12), (23)和(34)。

[20] 在此处我们不会区分因子分析和主成分分析,我们是在更广泛的意义上使用因子分析。

[21] 使用 RUMM2030、OPLM 或者 WINMIRA 软件得到的结果都是相同的。

参考文献

Andrich, D. (1988). Rasch models for measurement. *Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-068*. Newbury Park, CA: Sage.

Andrich, D., & Luo, G. (1993). A hyperbolic cosine latent trait model for unfolding dichotomous single-stimulus responses. *Applied Psychological Measurement, 17*, 253–276.

Bond, T. G., & Fox, C. M. (2007). Applying the Rasch model. *Fundamental measurement in the social sciences*. Mahwah, NJ: Erlbaum.

Browne, M. (1992). Circumplex models for correlation matrices. *Psychometrika, 57*, 469–497.

Carroll, J. B. (1945). The effect of difficulty and chance success on correlations between items or between tests. *Psychometrika, 10*, 1–19.

Coombs, C. H. (1950). Psychological scaling without a unit of measurement. *Psychological Review, 57*, 145–158.

Coombs, C. H. (1964). *A theory of data*. New York, NY: Wiley.

Coombs, C. H., Dawes, R. M., & Tversky, A. (1970). *Mathematical psychology: An elementary introduction*. Englewood Cliffs, NJ: Prentice Hall.

Embretson, S., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Ferguson, G. A. (1941). The factorial interpretation of test difficulty. *Psychometrika, 6*, 323–330.

Goodman, L. A., & Kruskal, W. H. (1979). *Measures of association for cross classifications*. New York, NY: Springer Verlag.

Guttman, L. (1950). The utility of scalogram analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Measurement and prediction. Studies in Social Psychology in World War II* (Vol. 4, pp. 122–171). New York, NY: Wiley.

Huisman, M. (1998). *Item nonresponse: Occurrence, causes, and imputation of missing answers to test items*. Leiden, The Netherlands: DSWO Press.

Huisman, M. (2000). Imputation of missing item responses: Some simple techniques. *Quality and Quantity, 34*, 331–351.

Jacoby, W. G. (1991). Data theory and dimensional analysis. *Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-078*. Newbury Park, CA: Sage.

Kingma, J., & Taerum, T. (1989). SPSS-X procedure and standalone programs for the Mokken scale analysis: A nonparametric item response theory model. *Educational & Psychological Measurement, 49*, 101–136.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 140*, 1–55.

Loevinger, J. (1948). The technique of homogeneous tests compared with some aspects of "scale analysis" and factor analysis. *Psychological Bulletin, 45*, 507–530.

McIver, J., & Carmines, E. G. (1981). Unidimensional scaling. *Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-078*. Newbury Park, CA: Sage.

Meijer, R. R. (1994). The number of Guttman errors as a simple and powerful person-fit statistic. *Applied Psychological Measurement, 18*, 311–314.

Mokken, R. J. (1971). *A theory and procedure of scale analysis. With applications in political research*. Berlin, Germany: De Gruyter (Mouton).

Mokken, R. J. (1997). Nonparametric models for dichotomous responses. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory.* New York, NY: Springer Verlag.

Mokken, R. J., & Lewis, C. (1982). A nonparametrical approach to the analysis of dichotomous item responses. *Applied Psychological Measurement, 6,* 417–430.

Mokken, R. J., Lewis, C., & Sijtsma, K. (1986). Rejoinder to "The Mokken scale: A critical discussion". *Applied Psychological Measurement, 10,* 279–285.

Mokken, R. J., van Schuur, W. H., & Leeferink, A. J. (2001). The circles of our minds. A nonparametric IRT model for the circumplex. In A. Boomsma, M. A. J. van Duijn, & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 339–356). New York, NY: Springer Verlag.

Molenaar, I. W. (1991). A weighted Loevinger H-coefficient extending Mokken scaling to multicategory items. *Kwantitatieve Methoden, 37,* 97–118.

Molenaar, I. W., & Sijtsma, K. (2000). User's manual MSP5 for Windows, A program for Mokken scale analysis for polytomous items, Version 5.0. Groningen, the Netherlands: Science Plus/iec ProGamma.

Molenaar, W. (1970). Approximations to the Poisson, binomial and hypergeometric distribution functions. MC Tract 31. Amsterdam, the Netherlands: Centrum voor Wiskunde en Informatica.

Niemöller, B., & van Schuur, W. H. (1983). Stochastic models for unidimensional scaling: Mokken and Rasch. In D. McKay, N. Schofield, & P. Whiteley (Eds.), *Data analysis and the social sciences* (pp. 120–170). London, England: Francis Pinter.

Ostini, R., & Nering, R. (2006). Polytomous item response theory models. *Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-144.* Thousand Oaks, CA: Sage.

Plutchik, R., & Conte, H. R. (Eds.). (1997). *Circumplex models of personality and emotions.* Washington, DC: American Psychological Association.

Popper, K. R. (1959). *The logic of scientific discovery.* London, England: Hutchinson.

Popper, K. R. (2003). Conjectures and refutations. *The growth of scientific knowledge.* London, England: Routledge. (Original work published 1963)

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests* (Copenhagen: Danish Institute for Educational Research), expanded edition. Chicago, IL: The University of Chicago Press. (Original work published 1960)

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2000). A general item response theory model for unfolding unidimensional polytomous responses. *Applied Psychological Measurement, 24,* 3–32.

Roskam, E. E., van den Wollenberg, A. L., & Jansen, P. G. W. (1986). The Mokken scale: A critical discussion. *Applied Psychological Measurement, 10,* 165–177.

Sanders, K., & Schuur, W. van. (1998). De Noorderlingen: Identiteit en Vertrouwen. *Sociale Wetenschappen, 41,* 24–48.

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory.* Thousand Oaks, CA: Sage.

Smith, E. V., & Stone, G. E. (2009). *Criterion referenced testing: Practice analysis to score reporting using Rasch measurement.* Maple Grove, MN: Jam Press.

Van der Ark, L. A. (2007). *Mokken scale analysis in R. Journal of Statistical Software, 20,* 1–19.

van Schuur, W. H. (1993). Nonparametric unidimensional unfolding for multicategory data. *Political Analysis, 4,* 41–74.

van Schuur, W. H. (2003). Mokken scale analysis: Between the Guttman scale and parametric item response theory. *Political Analysis, 11,* 139–163.

van Schuur, W. H. & Kiers, H. A. L. (1994). Why factor analysis often is the wrong model for analyzing bipolar concepts and what model to use instead. *Applied Psychological Measurement, 18,* 97–110.

Weesie, J. (1999). MOKKEN: Stata module: Mokken scale analysis. Software Components RePEc:boc:bocode:sjw31, RePEc EconPapers. http://econpapers.repec.org/software/bocbocode/sjw31.htm

Wood, R. (1978). Fitting the Rasch model: A heady tale. *British Journal of Mathematical and Statistical Psychology, 31,* 27–32.

进一步阅读书目

定序 IRT(按出版年份排序)

Molenaar, I. W. (1982). Mokken scaling revisited. *Kwantitatieve Methoden, 3,* 145–164.

Mokken, R. J., Lewis, C., & Sijtsma, K. (1986). Rejoinder to "The Mokken scale: A critical discussion." *Applied Psychological Measurement, 10,* 279–285.

Sijtsma, K., & Molenaar, I. W. (1987). Reliability of test scores in nonparametric item response theory. *Psychometrica, 52,* 79–97.

Sijtsma, K. (1988a). Reliability estimation in Mokken's nonparametric item response model. In W. E. Saris & I. N. Gallhofer (Eds.), *Sociometric research, Vol. I: Data collection and scaling* (pp. 159–174). London, England: MacMillan Press.

Sijtsma, K. (1988b). *Contributions to Mokken's nonparametric item response theory.* Amsterdam, The Netherlands: Free University Press.

Meijer, R. R., Sijtsma, K., & Smid, N. J. (1990). Theoretical and empirical comparison of the Mokken and the Rasch approach to IRT. *Applied Psychological Measurement, 14,* 283–298.

Sijtsma, K., Debets, P., & Molenaar, I. W. (1990). Using Mokken scale analysis to develop unidimensional scales. *Quality and Quantity, 24,* 173–188.

Hemker, B. T., Sijtsma, K., & Molenaar, I. W. (1995). Selection of unidimensional scales from a multidimensional item bank in the polytomous Mokken IRT model. *Applied Psychological Measurement, 19,* 337–352.

Sijtsma, K., & Junker, B. W. (1996). A survey of theory and methods of invariant item ordering. *British Journal of Mathematical and Statistical Psychology, 49,* 79–105.

Molenaar, I. W. (1997). Nonparametric models for polytomous responses. In W. J. Van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 369–380). New York, NY: Springer Verlag.

Sijtsma, K., & Hemker, B. T. (1998). Nonparametric polytomous IRT models for invariant item ordering, with results for parametric models. *Psychometrika, 63,* 183–200.

Van Abswoude, A. A. H., Vermunt, J. K., Hemker, B. T., & Van der Ark, L. A. (2004). *Mokken scale analysis using hierarchical clustering procedures. Applied Psychological Measurement, 28,* 332–354.

软件

Debets, P., Sijtsma, K., Brouwer, E., & Molenaar, I. W. (1988). MSP: A computer program for item analysis according to a nonparametric IRT approach. *Psychometrica, 54,* 534–536.

Kingma, J., & Reuvekamp, J. (1986). Mokken scale: A PASCAL program for nonparametric stochastic Mokken scales. *Educational and Psychological Measurement, 46,* 679–685.

Rivas, T., & Martinez, M. R. (1992). MOKPAS: Un programa para el escalamiento de items según el modelo TRI no paramétrico de Mokken. *Investigaciones Psicológicas,* 187–205.

Rivas, T., Martinez, M. R., & Hidalgo, R. (1996). MOKFOR1: A program to fit an accumulative scale to Mokken non parametric IRT model. 20th Biennial Conference of the Society for Multivariate Analysis in the Behavioral Sciences, ESADE, Barcelona.

Sijtsma, K., Debets, P., & Molenaar, I. W. (1990). Mokken scale analysis for polychotomous items: Theory, a computer program, and an empirical application. *Quality and Quantity, 24,* 173–188.

应用

Barker, M., O'Hanlon, A., McGee, H. M., Hickey, A., & Conroy, R. M. (2007). Cross-sectional validation of the Aging Perceptions Questionnaire: A multidimensional instrument for assessing self-perceptions of aging. BMC Geriatrics, 7, 9.

Bech, P., Hansen, H. V., & Kessing, L. V. (2006). The internalising and externalising dimensions of affective symptoms in depressed (unipolar) and bipolar patients. *Psychotherapy and Psychosomatics, 75,* 362–369.

Boor, K., Scheele, F., van der Vleuten, C. P., Scherpbier, A. J., Teunissen, P. W., & Sijtsma, K. (2007). *Psychometric properties of an instrument to measure the clinical learning environment. Medical Education, 41,* 92–99.

Cingranelli, D. L., & Richards, D. L. (1999). Measuring the level, pattern and sequence of government respect for physical integrity rights. *International Studies Quarterly, 43,* 407–417.

Davenport, C. (1995). Multidimensional threat perception and state repression: An inquiry into why states apply negative sanctions. *American Journal of Political Science, 39,* 683–713.

de Jong, A., & Molenaar, I. W. (1987). An application of Mokken's model for stochastic, cumulative scaling in psychiatric research. *Journal of Psychiatric Research, 21,* 137–149.

Duivenvoorden, H. J., Tibboel, D., Koot, H. M., van Dijk, M., & Peters, J. W. (2006). Pain assessment in profound cognitive impaired children using the Checklist Pain Behavior: Is item reduction valid? *Pain, 126,* 147–154.

Ettema, T. P., Dröes, R. M., de Lange, J., Mellenbergh, G. J., & Ribbe, M. W. (2007). QUALIDEM: Development and evaluation of a dementia specific quality of life instrument. Scalability, reliability and internal structure. *International Journal of Geriatric Psychiatry, 22,* 549–556.

Gillespie, M., Tenvergert, E. M., & Kingma, J. (1987). Using Mokken scale analysis to develop unidimensional scales. Do the six abortion items in the NORC GSS form one or two scales? *Quality and Quantity, 21,* 393–408.

Gillespie, M., Tenvergert, E. M., & Kingma, J. (1988). Using Mokken methods to develop robust cross-national scales: American and West German attitudes toward abortion. *Social Indicators Research, 20,* 181–203.

Hosenfeld, B., Van den Boom, D. C., & Resing, W. C. M. (1997). Constructing geometric analogies for the longitudinal testing of elementary school children. *Journal of Educational Measurement, 34,* 367–372.

Ivarsson, B., & Malm, U. (2007). Self-reported consumer satisfaction in mental health services: Validation of a self-rating version of the UKU-Consumer Satisfaction Rating Scale. *Nordic Journal of Psychiatry, 61,* 194–200.

Jacoby, W. G. (1994). Public attitudes towards government spending. American *Journal of Political Science, 38,* 336–361.

Jacoby, W. G. (1995). The structure of ideological thinking in the American electorate. *American Journal of Political Science, 39,* 314–335.

Kingma, J., & Reuvekamp, J. (1984). The construction of a developmental scale for seriation. *Educational and Psychological Measurement, 44,* 1–23.

Kingma, J., & Te Vergert, E. M. (1985). A nonparametric scale analysis for the development of conservation. *Applied Psychological Measurement, 9,* 375–387.

Koh, C. L., Hsueh, I. P., Wang, W. C., Sheu, C. F., Yu, T. Y., Wang, C. H., & Hsieh, C. L. (2006). Validation of the action research arm test using item response theory in patients after stroke. *Journal of Rehabilitation Medicine, 38,* 375–380.

Kørner, A., Lauritzen, L., Abelskov, K., Gulmann, N. C., Brodersen, A. M., Wedervang-Jensen, T., & Marie Kjeldgaard, K. (2007). Rating scales for depression in the elderly: External and internal validity. *Journal of Clinical Psychiatry, 68,* 384–389.

Lecrubier, Y., & Bech, P. (2007). The Ham D(6) is more homogenous and as sensitive as the Ham D(17). *European Psychiatry, 22,* 252–255.

Licht, R. W., Qvitzau, S., Allerup, P., & Bech, P. (2005). Validation of the Bech–Rafaelsen Melancholia Scale and the Hamilton Depression Scale in patients with major depression: Is the total score a valid measure of illness severity? *Acta Psychiatrica Scandinavica, 111,* 144–149.

Luinge, M. R., Post, W. J., Wit, H. P., & Goorhuis-Brouwer, S. M. (2006). The ordering of milestones in language development for children from 1 to 6 years of age. *Journal of Speech, Language and Hearing Research, 49,* 923–940.

Olsen, L. R., Mortensen, E. L., & Bech, P. (2004). The SCL-90 and SCL-90R versions validated by item response models in a Danish community sample. *Acta Psychiatrica Scandinavica, 110,* 225–229.

Paas, L. J. (1998). Mokken scaling characteristic sets and acquisition patterns of durable and financial products. *Journal of Economic Psychology, 19,* 353–376.

Paas, L. J., & Molenaar, I. W. (2005). Analysis of acquisition patterns: A theoretical and empirical evaluation of alternative methods. *International Journal of Research in Marketing, 22,* 87–100.

Roorda, L. D., Roebroeck, M. E., van Tilburg, T., Lankhorst, G. J., & Bouter, L. M., Measuring Mobility Study Group. (2004). Measuring activity limitations in climbing stairs: Development of a hierarchical scale for patients with lower-extremity disorders living at home. *Archives of Physical Medicine and Rehabilitation, 85,* 967–971.

Scarritt, J. R. (1996). Measuring political change: The quantity and effectiveness of electoral and party participation in the Zambian one-party state, 1973–1991. *British Journal of Political Science, 26,* 283–297.

Schneider, S. K., Jacoby, W. G., & Coggburn, J. D. (1997). The structure of bureaucratic decisions in the American states. *Public Administration Review, 57,* 240–249.

Segura, S. L., & Gonzalez-Roma, V. (2003). How do subjects construe ambiguous response formats of affect items? *Journal of Personality and Social Psychology, 85,* 956–968.

Sijtsma, K., Emons, W. H., Bouwmeester, S., Nyklícek, I., & Roorda, L. D. (2008). Nonparametric IRT analysis of quality-of-life scales and its application to the World Health Organization Quality-of-Life Scale (WHOQOL-Bref). *Quality of Life Research, 17,* 275–290.

Sijtsma, K., & Verwey, A. (1992). Mokken scale analysis: Theoretical considerations and an application to transitivity tasks. *Applied Measurement in Education, 5,* 355–373.

Stochl, J., Boomsma, A., van Duijn, M., Brozová, H., & Růzicka, E. (2008). Mokken scale analysis of the UPDRS: Dimensionality of the Motor Section revisited. 1. *Neuro Endocrinology Letters, 29,* 151–158.

Van der Putten, A., Vlaskamp, C., Reynders, K., & Nakken, H. (2005). Movement skill assessment in children with profound multiple disabilities: A psychometric analysis of the top down motor milestone test. *Clinical Rehabilitation, 19,* 635–643.

Van der Veer, K., Ommundsen, R., Larsen, K. S., Le, H. V., Krumov, K., Pernice, R. E., & Romans, G. P. (2004). Structure of attitudes toward illegal immigration: Development of cross-national cumulative scales. *Psychological Reports, 94,* 897–906.

van Schuur, W. H., & Vis, J. C. P. M. (2000). What Dutch parliamentary journalists know about politics. *Acta Politica, 35,* 196–227.

Verweij, A. C., Sijtsma, K., & Koops, W. (1996). A Mokken scale for transitive reasoning suited for longitudinal research. *International Journal of Behavioral Development, 19,* 219–238.

Watson, R., Deary, I. J., & Shipley, B. (2008). A hierarchy of distress: Mokken scaling of the GHQ-30. *Psychological Medicine, 38,* 575–579.

Zinn, F. D., Henderson, D. A., Nystuen, J. D., & Drake, W. D. (1992). A stochastic cumulative scaling method applied to measuring wealth in Indonesian villages. *Environment and Planning A, 24,* 1155–1166.

译名对照表

circular unfolding model	循环展开模型
circumplex	环状
classical test theory	经典测试理论
coefficient of homogeneity	同质性系数
coefficient of reproducibility	再现性系数
coefficient of scalability	可量测性系数
cumulative scales	累计量表
double monotonicity model	双重单调性模型
dominance relation	支配关系
dominance response	支配应答
error cell	误差单元格
fundamental measurement	基本测量
Guttman scale	哥特曼量表
implicational scale	蕴含量表
item-response theory(IRT)	题项回答理论
item-response function(IRF)	题项回答函数
item Step	题项梯度
latent	潜在
latent variable	潜变量
local stochastic independence	局部随机独立性
model violation	模型违反
monotonely homogeneous item	单调同质性题项
monotone homogeneity model	单调同质性模型
nondiscriminating response patterns	不具区分度的回答模式
ordinal circular unfolding model	定序循环展开模型
ordinal scale	定序量表
parametric IRT model	参数 IRT 模型
probabilistic dominance model	概率性支配模型
proximity relations	临近关系
proximity response	临近应答
Rasch model	拉希模型

reliability	信度
restscore group	余分分组
restsplit group	合并余分分组
scalogram	量表图示法
the coefficient of reproducibility	再现性系数
the coefficient of scalability	可量测性系数
the coefficient of homogeneity	同质性指数
transitivity relationships	传递性关系